THÉORIES ET TECHNIQUES DE BASE POUR L'ANALYSE DE DONNÉES EN PSYCHOLOGIE

Inférence - analyse de puissance - tests d'hypothèse
prédiction linéaire simple et multiple (régression)
structuration de données multiples

Année 2008

TABLE DES MATIÈRES

A. Introduction : qu'est-ce que « *les statistiques* » ?

Les statistiques, telles que nous les connaissons aujourd'hui, constituent un ensemble de théories et de techniques extrêmement variées, remplissant des tâches diverses dont les relations ne sont pas toujours claires ; description, estimation, test de modèles, prédiction et bien d'autres. Quelles sont en fait les relations entre ces différents objectifs et le calcul des probabilités, la pratique des tests d'hypothèse, le raisonnement inférentiel, l'analyse combinatoire, etc. ?

Lorsqu'un « profane » s'exprime sur « les statistiques », il recourt à l'une des plus anciennes conceptions des statistiques, à savoir celle d'un ensemble de techniques de calcul plus ou moins indigestes visant à décrire l'état présent d'une collectivité ou d'un quelconque groupe, humain ou non. La *Staatistik* est née pendant la première moitié du XIXe siècle avec l'introduction de recensements et fut parfois considérée par les esprits les plus réformateurs de l'époque comme « *la vraie science d'état* ». Cette science nouvelle a une vocation essentiellement *descriptive*, son objectif est de décrire des *faits*, c'est-à-dire de *compter des fréquences* et des pourcentages, éventuellement de *calculer des moyennes et des écart-types*. Depuis quelques décennies, et surtout depuis le développement fulgurant des moyens informatiques, on range également dans les techniques statistiques descriptives l'analyse factorielle (ACP et analyse de correspondances), ainsi que toutes les techniques dérivées de l'analyse canonique (analyse discriminante). On parle dans ces cas d'*analyse exploratoire*.

La statistique probabiliste a des visées beaucoup plus générales : il ne s'agit pas seulement de décrire une réalité limitée à des circonstances données, mais d'imaginer un modèle théorique dont cette réalité observée fortuitement n'est qu'une « *réalisation* » (au sens statistique : découlant d'une expérience aléatoire) parmi d'autres. Dans cette optique, la tâche de la recherche est certes de décrire certaines observations, mais aussi de tester l'adéquation d'un certain modèle à ces observations. Depuis le début de ce siècle, un arsenal impressionnant de tests d'hypothèse a été développé à cette fin. Dans les cas multivariés, les analyses factorielles confirmatoires jouent le même rôle : il s'agit de tester l'adéquation d'une structure théorique à une structure observée. En plus de la simple description, le second rôle des statistiques est donc d'ordre décisionnel : elles permettent, dans certaines limites de confiance, de *décider* si oui – ou non – une certaine régularité observée localement peut être généralisée à un ensemble plus général, à savoir la population.

« Les statistiques » remplissent encore un troisième rôle qui consiste à *réaliser l'inférence* : c'est-à-dire *estimer* et *prédire*. Ces deux termes ne sont pas superposables. Supposons par exemple que le lien (déclaré non nul par un test d'hypothèse convenable) entre deux grandeurs mesurées X et Y sur un échantillon puisse être considéré comme linéaire et que l'équation les liant s'exprime sous la forme $aX + b = Y$. Les paramètres a et b de l'équation ci-dessus ne sont que des estimations des paramètres théoriques

inconnus α et β, de l'équation « théorique » valable pour toute la population ($\alpha X + \beta = Y$). L'*estimation statistique* permet de trouver, étant donné des circonstances expérimentales déterminées, les meilleures estimations possibles a et b de α et β. Dans un second temps, l'estimation d'un modèle ayant été réalisée, on peut utiliser l'équation « incarnant » le modèle pour prédire tout score Y, connaissant X. Il ne s'agit donc plus d'estimation à proprement parler, mais d'une *prédiction statistique* réalisée à l'aide d'un modèle qui, lui, est estimé.

A.1. **Décrire, estimer et prédire :** deux exemples *théoriques* :

- *« Pile ou Face » ?*

Le jet d'une pièce de monnaie constitue l'expérience aléatoire la plus simple et la mieux connue que l'on puisse réaliser, elle est à la base de tout le raisonnement probabiliste. Les trois tâches de la statistique peuvent s'appliquer à ce type d'expérience que l'on pourrait qualifier d'« archétypique ». Lançons une pièce 50 fois en l'air et notons les résultats :

- Tâche 1 des statistiques (Staatistik) ; compter les occurrences de P et de F, dresser un tableau, calculer des pourcentages.

- Tâche 2 des statistiques : on confronte ce résultat à des attentes, on se sert des données pour évaluer une hypothèse, par exemple que la pièce est vraie (équilibrée). Si l'on s'étonne que le résultat s'écarte d'une certaine attente, cela signifie qu'un « test » implicite a été opéré. A ce stade, il est primordial d'expliciter les modèles, de manière à pouvoir les formaliser. Le test d'hypothèse classique devient alors possible : la pièce est-elle équilibrée ou non ?

- Tâche 3 des statistiques : la recherche ne s'arrête pas avec l'affirmation selon laquelle la pièce est éventuellement truquée ; si c'est le cas, l'estimation statistique doit quantifier le déséquilibre des chances, c'est-à-dire construire un nouveau modèle tenant compte des données fournies par l'expérimentation. Dans cet exemple, l'expérience des 50 lancers permettra d'*estimer* les paramètres de la distribution des probabilités d'une certaine hypothèse, par exemple que la pièce est truquée dans le sens d'avoir 65% de Pile[1]. Ce calcul étant fait, il devient possible de *prédire* la répartition d'une série de nouveaux lancers.

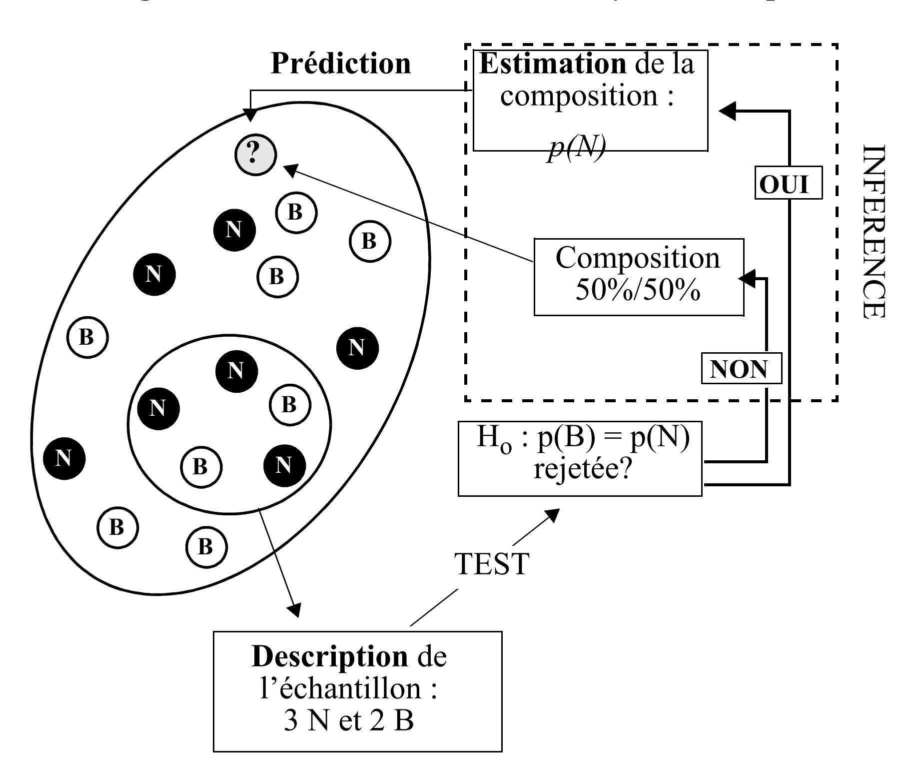
- *L'urne de Bernouilli*

Ce problème de pièce de monnaie ne semble pas avoir, à première vue, de rapport direct avec l'évaluation psychologique. Il constitue pourtant son modèle théorique fondamental, en particulier pour les items dont les réponses sont du type « vrai/faux ». En effet, chaque personne prise au hasard et répondant à un tel item peut être représentée au

1. On cherche en fait à décrire une distribution de probabilités d'une probabilité !

niveau formel comme un jet de pièce, et sa réponse comme l'observation d'un de ses côtés. Afin de mieux comprendre cette analogie, considérons un modèle intermédiaire : l'urne de Bernouilli. Considérons cette urne remplie de boules blanches et noires, dans une proportion inconnue, et tirons par exemple 50 boules.

Figure 1 : les trois niveaux de l'analyse statistique



- En ce qui concerne leur tâche *descriptive*, les statistiques permettent de connaître le nombre de boules de chaque couleur extraites de l'urne.

- De leur côté, les statistiques probabilistes, à vocation *inférentielle*, se posent la question de la composition de l'urne. Celle-ci est considérée comme une population (de taille infinie, si le tirage s'effectue avec remise), et le n-tirage (aléatoire) en constitue un échantillon. La question est de savoir ce qu'on peut « parier » à propos de la composition de l'urne, sur la base des résultats de ce n-tirage. On peut par exemple tester l'hypothèse que sa composition est de 50% – 50%, on se retrouve alors dans le cas de la pièce de monnaie. Si cette hypothèse devait être rejetée, de nouvelles questions se posent :

- L'**estimation statistique** va tenter de formuler un nouveau modèle de la composition de l'urne, sur la base des résultats du n-tirage, par exemple p(Blanche) = .40 ; cette estimation est bien entendu assortie elle-même d'une distribution de probabilités, donc d'un intervalle de confiance. Dans un tel cas, la prédiction statistique permet de « parier » sur un nombre de boules blanches proche de 40, pour un nouveau tirage de 100 boules, par exemple.

La stratégie de recherche décrite dans la figure ci-dessus (p. précédente) est celle qui a été adoptée généralement en psychologie.

A.2. Décrire, estimer et prédire : deux exemples empruntés aux sciences humaines :

- *Exemple 1*

Dans le cadre de la psychologie des intérêts, on peut se demander si les études en sciences humaines attirent autant les hommes que les femmes. Si on tire un échantillon d'effectif 100, la simple étude descriptive consiste à compter les étudiants des deux sexes. Si ce résultat devait permettre une inférence à la population globale des personnes susceptibles d'être attirées par ce type d'études, on peut, dans un premier temps, tester l'hypothèse que la répartition est équivalente. Si cette hypothèse devait être rejetée en regard des données, on peut estimer un modèle de répartition différent qui permettra de prédire la répartition en hommes et femmes d'une future volée d'étudiants en sciences humaines, ou la probabilité qu'une personne étudiant en Faculté des sciences humaines soit une femme ou un homme.

Application numérique : Un groupe d'étudiantes et d'étudiants (par exemple n = 100) en psychologie est constitué à 80% de femmes. A quelles conditions, et dans quelle « mesure », peut-on induire de cette observation que les femmes sont davantage intéressées par cette branche que les hommes ?

Traitement du problème :

- Il faut tout d'abord se demander quel est le modèle mis en cause par cette question. Il s'agit à l'évidence du modèle équiprobable, car se demander s'il y a une différence d'intérêt entre hommes et femmes pour la psychologie, c'est avant tout mettre en question une *Hypothèse nulle*, à savoir p(H) = p(F) = .5.
- Remarquons que nous traitons le problème d'une différence d'intérêt par le biais d'une observation portant sur des taux d'inscription aux filières qui nous intéressent. Ce lien n'est pas évident !
- Étant donné qu'il n'est pas possible d'observer tous les étudiants inscrits en psychologie dans tout le pays ou dans le monde en général, on doit se contenter d'observations portant sur un échantillon, en l'occurrence 100 personnes dont on espère qu'elles sont représentatives de tous les étudiants inscrits dans cette branche.
- On sait (…) que si un événement E (ici : être inscrit en psychologie *et* être une femme) se produit avec une fréquence de p dans une population, alors la distribution échantillonnale de la variable F (dont les scores sont des f_i, probabilités d'observer l'événement E dans un échantillon de 100 étudiants en psychologie) a les caractéristiques suivantes :
  - la variable F est gaussienne et a une espérance égale à p

- et un écart-type égal à : $s = \sqrt{\frac{(1-p) \cdot p}{n}}$

Avec n=100, taille de l'échantillon.

- Ces valeurs nous permettent de calculer la largeur d'un intervalle de confiance à 95%, par exemple, qui aura 95 chances sur 100 de contenir une valeur f observée au cours d'une expérience (celle que nous traitons ici), si H_o est vraie.
- Si notre valeur empirique, à savoir .8, n'est pas incluse dans l'intervalle de confiance calculé autour de .5, alors nous rejetterons l'hypothèse nulle avec 5% de chances de prendre une décision erronée.
- Calculons maintenant cet intervalle, dont la largeur vaut approximativement 2 écart-types de F, plus exactement :

$$1,96 \cdot s = 1,96 \cdot \sqrt{\frac{(1-(0,5)) \cdot 0,5}{100}} = 1,96 \cdot 0,05 = 0,1$$

- L'intervalle de confiance symétrique à 95% autour de .5 est donc borné par .40 et .60, et notre valeur empirique (.80) ne s'y trouve pas.
- On pourrait se demander quelle est la probabilité d'observer une fréquence f = .8 sous H_o. Cette valeur s'écarte de .80 – .50 = .30 de l'espérance p de la variable F, sous H_o. Cette valeur de .30 équivaut à .30/s = .30/.05 = 6 écart-types de l'espérance, ce qui rend cet événement extrêmement rare sous H_o.
- Nous voici donc amenés à rejeter le modèle d'équiprobabilité, ce qui signifie que si nous tirons au hasard une personne étudiant en psychologie, la probabilité de tirer une personne de sexe féminin est supérieure à celle de tirer une personne de sexe masculin.
- Le risque de se tromper en affirmant cela est de 5 pour cent, ce qui – pratiquement – signifie que sur 100 tirages d'échantillons de 100 personnes, 5 d'entre-eux fourniraient des valeurs f_i situées hors de l'intervalle de confiance autour de .5, *alors même que H_o serait vraie* !
- Le modèle *a priori* d'équiprobabilité étant rejeté, on adhérera provisoirement à un modèle *a posteriori*, à savoir celui d'une répartition de 80/20 %, mais ce nouveau modèle va devoir être confirmé par de nouvelles expériences testant l'hypothèse « nulle » : p(F) = .8. Et ainsi progresse la connaissance…

• *Exemple 2*

L'administration d'un canton s'interroge au sujet de l'opportunité de financer un programme de formation à la recherche d'emploi destiné à des apprentis. Le « traitement » consiste à proposer divers cours permettant aux apprentis de mieux pratiquer les différentes techniques de recherche d'emploi (rédiger un CV, une lettre, soutenir un entretien, téléphoner). L'indice quantitatif, ou « critère », utilisé pour mesurer

l'effet du cours est par exemple le degré de connaissance des techniques, auto-évalué par les apprentis.

Le chercheur assumant cette recherche tire (en principe…) un échantillon de chômeurs au hasard, afin que celui-ci soit représentatif de l'ensemble des personnes concernées. Le critère « connaissance des techniques » est mesuré deux fois, une fois avant le cours, une fois après. L'échelle utilisée est du type Likert, en six points. Les deux moyennes sont calculées, puis leur différence. On peut aussi détailler les résultats, élève par élève, tout en précisant leur sexe, leur âge, niveau scolaire, classe, établissement, commune, etc. Cette phase correspond à l'*étape descriptive* qui aboutit en général à des tableaux dans lesquels figurent des effectifs, des pourcentages, des moyennes et des écart-types. L'interprétation de ces résultats peut révéler que les différences *après* - *avant* sont positives, donc que le cours aurait entraîné un certain progrès dans la connaissance des techniques, *pour le groupe considéré*.

C'est l'objet d'une seconde étape que de convaincre le commanditaire de la recherche que ces résultats sont inférables à la population des apprentis en général. On proposera donc certains modèles permettant de comprendre la réaction *générale* de tout apprenti à ce type de « traitement ». Curieusement (mais très classiquement), le modèle le plus souvent testé est celui de l'inefficacité absolue du traitement (H_0), c'est-à-dire le modèle de l'*indépendance* entre les variables *moyenne au critère* et la *variable catégorielle Avant/Après*. Le choix de ce type de modèle implique évidemment un fort désir de rejeter H_0 , puisque la décision de rejet signifierait que le traitement est efficace, pour tout apprenti de la population considérée. Le chercheur espère donc observer des niveaux de signification petits, inférieurs au seuil conventionnel de 5%. Mais cette technique est de maigre rendement : elle ne dit pas grand chose sur l'ampleur de l'éventuel progrès mis en évidence par le test. Or, répondre à cette question en exhibant des p-values faibles est incorrect, car leur valeur dépend de la taille de l'échantillon. En effet, pour un effectif très grand, la plus infime différence entre moyennes peut être déclarée « très significative » par un test de Student (utilisé dans le cas de cet exemple). La signification statistique n'informe donc pas sur l'*intérêt* ou la *signification psychologique* de l'effet observé. Elle se borne à déclarer que l'effet observé n'est pas nul - et qu'il n'est pas possible de l'attribuer au seul hasard de l'échantillonnage.

Or, pour le commanditaire de la recherche, la question fondamentale est la suivante : quel progrès minimum (ou moyen) le cours peut-il garantir pour justifier son financement ?

Il est beaucoup plus pertinent dans ce cas de postuler des hypothèses plus spécifiées que celle de l'indépendance, par exemple : H_0 : « le cours augmente la moyenne du critère de 1 point sur une échelle de 6 ». Raisonnablement, on s'attendrait à ce que cette hypothèse soit discutée avec le commanditaire avant le début de toute expérimentation (c'est rarement le cas !).

Plus pertinente encore est l'*estimation* d'un modèle de l'effet du cours sur le critère : mais ce type de calcul fait appel à des techniques plus complexes (statistiques bayésiennes) qu'il est encore rare de rencontrer dans le domaine des sciences humaines.

Dans un troisième temps, les modèles postulés (ayant résisté aux tests) ou ceux qui auraient éventuellement été estimés, peuvent être affinés en fonction de divers paramètres : il se peut que le cours soit plus utile pour certaines techniques que pour d'autres, ou qu'il profite mieux à certains groupes qu'à d'autres (âge, sexe, niveau, etc.).

Finalement, grâce à ces modèles, on pourra *prédire* le niveau de connaissance des élèves après le cours, sur la seule base de leurs réponses « avant ». Ceux qui ont des résultats prédits trop faibles par rapport à une norme pourraient bénéficier de cours spéciaux ou d'un encadrement mieux personnalisé, par exemple.

Une dernière remarque s'impose : l'observation de notables améliorations dans la connaissance des techniques de recherche d'emploi ne s'explique pas nécessairement par le seul effet de la formation : on peut aussi supposer que le fait d'interroger les apprentis au sujet de leurs connaissances (avant toute forme d'intervention) ait suffi à les intéresser au problème, et suscité des discussions fructueuses avec leurs parents, amis, etc. Du point de vue purement expérimental, le plan décrit ci-dessus ne permet donc pas de déclarer que le traitement est utile ! Une recherche plus rigoureuse aurait exigé un plan plus complexe comportant au moins un groupe témoin, testé deux fois, mais n'ayant pas suivi le cours. Il semble que pratiquement, cette exigence soit le plus souvent impossible à satisfaire...

B. Les bases du raisonnement statistique moderne

Après avoir montré les différents objectifs des techniques statistiques appliquées dans la recherche en sciences humaines, il est temps de présenter les principaux courants de pensée qui ont contribué à la conception de la théorie statistique moderne, dans le contexte historique de leur développement.

B. 1. L'origine de la pensée probabiliste formalisée

On fait volontiers remonter les origines de la pensée scientifique à l'Antiquité grecque mais c'est à Kepler le premier, et surtout à Newton et Leibniz quelques années plus tard que revient traditionnellement la paternité des premiers développements scientifiques formalisés décrivant des « lois naturelles » (*régularités*) exprimées sous la forme d'équations mathématiques. Cependant, la brillante tradition scientifique qui s'en suivit dut attendre le début du 18e siècle pour être enfin en mesure de maîtriser un problème qui embarrassait tous les expérimentalistes et observateurs de la nature, aussi bien ceux attachés à l'étude de l'infiniment grand que de l'infiniment petit : il s'agit du problème des *erreurs* de mesure.

C'est probablement à *Gauss* et à *Laplace* que le contrôle des erreurs de mesure a pu devenir possible, et permettre enfin l'éclosion des méthodes d'analyse de données « modernes », encore utilisées de nos jours en physique, ainsi qu'en sciences naturelles et humaines (et ailleurs encore). C'est en effet à Gauss (vénéré aux Etats Unis comme l'un des plus grands génies scientifiques de tous les temps) qu'on attribue la première utilisation (vers 1810) d'un outil probabiliste (la loi dite *normale*) au service de l'inférence sur la mesure vraie d'une distance en astronomie. L'apport de Gauss à la pensée scientifique fut de réunir trois courants de la pensée scientifico-philosophique, ayant suivi depuis leur naissance, environ un siècle plus tôt, des évolutions parallèles :

- L'approche scientifique classique, héritée de Newton et des grands astronomes du 18e siècle, cherchait à tirer des lois générales à partir d'observations aussi précises que possible (astronomie, applications fondamentales à la navigation maritime, par exemple). Rappelons l'intuition géniale de Newton, caricaturée par l'événement de la chute de la pomme : le grand Newton peut « expliquer » l'événement (le *mouvement* de la pomme par rapport à la terre), en toute généralité, par une équation mathématique simple faisant intervenir la masse des objets en présence (la terre et la pomme), ainsi que le carré de leur distance.

- *L'approche inférentielle*, d'essence plus logico-philosophique, ne s'intéressait pas aux mouvements des corps physiques. Son objectif était de formaliser – en recourant également aux mathématiques – les mécanismes de la pensée inductive classique. En introduisant la notion de *probabilité conditionnelle* et celle de probabilité « *a priori* », Bayes (1702 - 1761) montra le premier la voie qui conduira à la possibilité d'estimer les paramètres d'un modèle abstrait, destiné à décrire une réalité inobservable directement.

- *L'approche statistique probabiliste*, d'essence purement mathématique va dériver de l'analyse combinatoire les principaux outils probabilistes utilisés en physique et dans la psychologie différentielle. La loi normale, ou « loi de fréquence des erreurs » reste la création statistique probabiliste la plus célèbre, elle servira de modèle au traitement des erreurs de mesure en physique, puis dans la plupart des disciplines ayant recours à des mesures. Toutes les autres lois de distribution théoriques (chi carré, Student, F) et toutes les lois multivariées) seront dérivées de ce modèle unique dont la conception mérite quelques éclaircissements.

On peut supposer que le modèle probabiliste de l'erreur intéressa très vite les physiciens qui s'empressèrent de l'adopter dès que Laplace et Gauss en fournirent l'équation exacte. Par contre, l'inférence et la pensée inductive n'intéressait guère les astronomes et physiciens, héritiers de Newton. N'ayant pas à travailler sur des échantillons, mais sur une réalité directement observable (masses, distances, temps), les scientifiques de l'époque pouvaient limiter leur travail à la recherche et à la vérification des régularités (lois) qu'ils s'efforçaient de traduire en équations. Ce qui était vrai pour tel phénomène physique l'était bien entendu pour tous les autres phénomènes semblables, faisant intervenir des objets semblables dans des circonstances semblables. Par exemple,

l'étude du fonctionnement d'un seul rein permet une connaissance du rôle de cet organe dans l'organisme humain *en général*.

Cependant, si la physique newtonienne ne fait pas référence explicitement à la logique inférentielle, c'est pourtant bien sur des *estimations* qu'elle base ses calculs lorsqu'elle prend pour valeur « vraie » d'une mesure la *moyenne* de toutes les mesures effectuées, considérées comme entachées d'erreurs. C'est donc lorsque la physique, et en l'occurrence l'astronomie, se préoccupa de décrire la répartition des erreurs autour d'une valeur hypothétique considérée comme *vraie*, qu'elle intégra le premier modèle *clef en main* fourni par l'approche statistique probabiliste. Après l'astronomie, ce fut la *thermodynamique* qui intégra le plus efficacement l'outil probabiliste. En 1857, Clausius (« *The nature of the motion which we call Heat* ») jette les bases de la physique statistique, bientôt suivi par Maxwell (1860) et Boltzmann, le fondateur de la physique statistique moderne.

Le modèle normal de l'erreur et l'inférence sur la moyenne seront exploités en sciences humaines dès la moitié du 19e siècle par les sociologues (Quételet) puis par les premiers psychologues différentialistes, dont le plus célèbre reste Galton, dont il sera question plus loin à propos de la découverte du phénomène de « régression ».

En résumé, nous retiendrons que la mise au point des premiers outils probabilistes applicables aux sciences de la nature date donc du début du 19e siècle. Ils furent utilisés principalement en physique et en astronomie, mais aussi en sociologie avec Quételet, puis en psychologie vers la fin du siècle, avec Galton et son étrange découverte de la « réversion », puis avec les premiers écrits de *Spearman* sur la construction d'échelles d'aptitudes (1904).

Curieusement, le développement des techniques inférentielles associées (tests de normalité, etc.) dut attendre les travaux du mathématicien anglais K. *Pearson*, qui fut le premier (vers 1898 seulement[2]) à mettre au point des *tests d'ajustement*, à l'intention des astronomes et généticiens. Ce n'est finalement que dans les années 1920 - 1930 que la *biométrie* (*Fisher* écrit « *statistical methods for research workers* » en 1925) et la *psychologie* (*Spearman* expose sa conception factorielle des aptitudes humaines en 1926) intègrent le raisonnement statistique inférentiel en l'appliquant aux modèles de mesure utilisés en psychologie[3]. La théorie des tests fournit un bel exemple de cette association dans la théorie classique de construction des tests, appelée précisément : *théorie de l'échantillonnage du domaine*.

2. À la décharge de toutes les personnes résistantes ou imperméables au mode de raisonnement statistique, remarquons que ces théories sont nées très tardivement dans l'histoire de la pensée scientifique et que, de plus, leur développement fut, comme nous l'avons vu, plutôt lent.

3. Remarquons que cette période fut une des plus riches de la physique, puisqu'elle vit le développement de la théorie de la relativité (Einstein) et celle des quanta, avec M. Planck. W. Heisenberg résuma les apports de la mécanique ondulatoire, fondamentalement probabiliste, en 1926.

B. 2. Les principaux outils probabilistes utilisés en psychologie

Développés et appliqués dès le 19e siècle et réellement popularisés après la seconde guerre mondiale, les outils probabilistes utilisés en psychologie tirent cependant leur origine 250 ans plus tôt, dans les méditations de quelques passionnés de jeu du 17e siècle. L'*analyse combinatoire* doit sa naissance, vers 1650, à la rencontre et à l'amitié de trois personnages fort différents, mais fortuitement intéressés par la même problématique, à savoir celle de la probabilité. *B. Pascal*, génie universel, esprit religieux, préoccupé par le problème de l'existence de Dieu, rencontre un personnage de cour, le chevalier de Méré, soucieux de maximiser ses gains au jeu de hasard. Pascal soumet ce problème à son ami mathématicien *Fermat*. Une riche correspondance s'en suivit, aboutissant – entre autres – à l'invention du triangle de Pascal, base de l'analyse combinatoire et premier jalon de la découverte de la loi binomiale par Newton, puis de la loi normale[4].

- *La loi normale*

Depuis son invention par Laplace et Gauss, la loi normale a joui d'une popularité grandissante et rien ne semble aujourd'hui encore pouvoir mettre son règne en péril. Il faut pourtant savoir qu'au début de ce siècle déjà, le mathématicien français Poincaré ironisait à son sujet : « *Tout le monde y croit [...] car les expérimentalistes s'imaginent que c'est un théorème mathématique, et les mathématiciens que c'est un fait expérimental* ».

La glorieuse histoire de la loi normale commence avec une intuition du philosophe et scientifique Blaise Pascal qui semble avoir été le premier à avoir suggéré l'existence d'un lien formalisé entre une équation mathématique et une série d'événements déterminés par le hasard. Il découvrit que les développements du binôme de Newton, arrangés sous la forme d'un triangle, donnaient exactement la description quantitative des différentes combinaisons d'occurrences Pile ou Face au jeu de la pièce de monnaie. En effet…

Si l'on jette une pièce *deux* fois, on a la possibilité d'observer *trois* types de combinaisons de P et de F, auxquelles correspondent certaines fréquences bien précises :

- celles comportant deux P :	PP	1
- celles comportant un seul P :	PF FP	2
- et celles où P est absent :	FF	1

Par exemple, la probabilité d'avoir un seul P en deux lancers est donc de 2/4 = .5. Et le plus curieux pour les esprits de l'époque fut de constater que les occurrences de ces combinaisons correspondaient exactement aux termes numériques du développement du binôme de Newton, que tout le monde apprend encore aujourd'hui à l'école : $(a+b)^2 = \mathbf{1} \cdot a^2 + \mathbf{2} \cdot ab + \mathbf{1} \cdot b^2$.

4. L'analyse combinatoire et le calcul des probabilités sont indissociables : si l'on veut par exemple connaître la probabilité d'observer un total de 10 en jouant deux dés, il faut connaître le nombre de combinaisons donnant un total de 10, en le rapportant à toutes les combinaisons possibles.

Si l'on jette une pièce *trois* fois, on a la possibilité d'observer *quatre* types de combinaisons de P et de F, auxquelles correspondent à nouveau certaines fréquences bien précises :

- celles comportant trois P :	PPP	1
- celles comportant deux P :	PPF PFP FPP	3
- celles comportant un P :	PFF FPF FFP	3
- et celles où P est absent :	FFF	1

Par exemple, la probabilité d'avoir un seul P en trois lancers est donc de 3/8 = .375. Et on retrouve les termes du binôme $(a + b)^3$.

Et ainsi de suite, en progressant dans les étages du triangle de Pascal et en dessinant un graphe pour chaque ligne, on voit peu à peu se dessiner l'allure caractéristique de la loi divine…

En 1657, le mathématicien hollandais C. Huygens s'intéresse passionnément à ces problèmes et vient à Paris pour s'initier à ces nouvelles théories. Déçu par la discrétion de Fermat qui le prit peut-être uniquement pour un passionné de gain, il retourne en Hollande pour écrire… un traité sur l'art de calculer les gains aux jeux de hasard (*De Ratiocinis Ludo Alea*). Ses écrits sont lus par Jacob (James) Bernouilli qui fonde la théorie de l'analyse combinatoire dans son *Ars Conjecturandi* (1713). C'est aussi à J. Bernouilli que l'on doit la première loi proprement statistique, connue aujourd'hui sous le nom de « *loi des grands nombres* »[5].

Cent ans après Pascal, le mathématicien *De Moivre* généralise la loi binomiale au cas continu, ce qui revient à faire tendre n vers l'infini dans la formule : $(a + b)^n$. La « *loi de fréquence des erreurs* » est conceptuellement prête mais ne trouve pas encore d'applications pratiques. Ce n'est que grâce aux perfectionnements permis par le calcul différentiel et intégral dû à *Leibniz*, qu'elle trouvera, chez Gauss et Laplace (indépendamment, semble-t-il) la formulation mathématique qui est encore utilisée de nos jours. L'Allemand *Gauss* l'appliquera pour la première fois en astronomie en 1810 ; de son côté, en France, *Laplace* en donne la formulation moderne dans sa *Théorie Analytique des Probabilités*, ouvrage dans lequel il démontre, entre autres, le théorème fondamental de l'inférence statistique, le « *théorème central limite* »[6].

5. Cette loi définit pour la première fois dans l'histoire des sciences une relation entre l'observation d'un événement particulier et celle portant sur une série de réalisations semblables, mais effectuées « au long cours ». La loi des grands nombres dit que si un événement unique se produit avec une probabilité p, alors la fréquence moyenne de cet événement, lors d'expérimentations répétées en nombre n, tend vers p lorsque n devient très grand. Par exemple, un très grand nombre de lancers d'une pièce équilibrée donneront des taux de *Pile* et de *Face* très proches. À la limite, si on lançait la pièce une infinité de fois, le rapport des taux serait de 50% - 50% exactement.
À noter que cette loi est à l'origine d'une croyance erronée qui veut que si on obtient une longue série de *Pile* consécutifs, alors la probabilité d'obtenir *Face* au coup suivant serait supérieure, comme s'il s'agissait de « compenser » la moyenne postulée par la loi des grand nombres. Cette intuition est trompeuse car les lancers successifs sont des événements indépendants.

La *loi normale*, ou *loi de Laplace-Gauss*, a été appliquée de diverses manières en psychométrie, donnant lieu à différentes interprétations de ses paramètres.

- *Dans une perspective purement descriptive*, lorsque la loi normale s'applique à la distribution des scores d'une population ou d'un échantillon, généralement des êtres humains, la variabilité des scores autour de leur moyenne, doit être interprétée ici comme une *dispersion* mesurée par l'écart-type (ou la variance). Si l'on est en présence de résultats à un test, cette dispersion mesure la *discrimination* du test, c'est-à-dire son « pouvoir séparateur », autrement dit son utilité.
  Dans ce cas, le score moyen n'a pas d'interprétation particulière, à moins de croire à l'existence de l'« homme moyen » de Quételet, auquel cas il correspondrait à la mesure idéale de l'homme-type voulu par Dieu.

- *Dans une perspective inférentielle*, la loi normale permet de décrire la dispersion d'un estimateur, par exemple la moyenne des moyennes de plusieurs échantillons de même taille, tirés d'une même population. Dans ce cas, la dispersion de la variable aléatoire « moyenne » (encore un modèle mathématique) est appelée *erreur d'échantillonnage*. C'est ce type d'erreur qui est évalué et analysé dans les procédures du type « t de Student » ou « analyse de variance ».

- Lorsqu'on applique la loi normale à la dispersion des scores prédits possibles, correspondant à un seul score « prédicteur », grâce à un modèle de régression estimé, on est en présence d'une *erreur d'estimation* (ou de prédiction).
  Indirectement, cette erreur est aussi due à l'échantillonnage puisque celui-ci conduit à calculer un modèle de prédiction estimé, et non théorique (auquel cas on ne parle plus d'erreur, mais de *résidu*).

- Enfin, lorsque la loi normale s'applique au score brut individuel pour décrire la distribution de tous les scores qu'un seul individu aurait obtenus au même test dans toutes les situations possibles, on parle d'*erreur de mesure* et on se place au même niveau d'interprétation qu'un physicien face à l'incertitude de sa mesure. Le traitement de ce type d'erreur a été abordé au cours « évaluation psychologique ».

Notons que lorsque l'on veut construire un bon test, il est nécessaire de répéter au moins deux conditions (en plus de la validité) : il faut que les scores soient précis (bonne fidélité) et que la discrimination des individus soit aussi bonne que possible. En termes statistiques, il faut que la dispersion des scores soit aussi large que possible *et* que l'erreur sur chaque score soit, pour sa part, aussi petite que possible. Le concepteur de tests se

6. Le théorème central limite est considéré comme le fondement de l'inférence statistique car il permet l'estimation d'un paramètre inconnu (valable au niveau général d'une population), sur la base d'un paramètre empirique mesuré sur une partie limitée, accessible, de cette population (l'échantillon).
Le Théorème de la Limite Centrale (ainsi mieux nommé par Saporta, 1990) affirme que si un certain caractère mesuré sur une population a une moyenne μ (généralement inconnue), alors la moyenne des moyennes de tous les échantillons de même taille tirés de cette population est la meilleure estimation de μ. Le théorème est encore plus fort car il permet aussi une estimation de la variance : la variance de la distribution des moyennes de tous les échantillons de taille n est la meilleure estimation de σ^2/n (σ^2 étant la variance du caractère dans la population).

trouve donc face à ce qui a été parfois appelé le *paradoxe psychométrique* : utilisant le même modèle appliqué au même objet, il doit travailler simultanément sur deux plans d'interprétation différents. Sur le plan de la distribution des scores de la population, il doit maximiser la dispersion en ajustant le temps de passation par exemple, ainsi qu'en sélectionnant les items dont la corrélation avec le total est suffisamment élevée. Sur le plan de la précision du score individuel, il doit minimiser l'erreur (augmenter la fidélité) en jouant sur le nombre d'items, c'est-à-dire en les présentant en nombre suffisant pour satisfaire à certains critères mathématiques (KR_{20} de Kuder-Richardson, formule de *Spearman* Brown), mais en veillant à ne pas introduire certains items qui diminueraient la qualité de la dispersion. Cette opération délicate porte le nom d'analyse d'items : elle repose en grande partie sur des critères empiriques et ne peut être formalisée de manière rigide.

- *La moyenne et le modèle normal de l'erreur*

La notion de *moyenne* a été introduite par le physicien T. Simpson (*An attempt to show the advantage arising by taking the mean of a number of observations in astronomy*, Philosophical transactions, 1755), A. Quételet, Statisticien d'État belge, reprit cette idée et l'appliqua à la description des populations. C'est lui qui formula l'hypothèse de l'« homme moyen », prototype idéal de l'homme « parfait » tel que voulu par le créateur, dont les humains réels ne sont que des avatars plus ou moins bien réussis (*cf.* Desrosières, p. 98). Quételet connaissait aussi la loi normale de Gauss et Laplace et sa conception d'une humanité globalement diverse, dont les caractéristiques fluctuent *normalement* autour d'une valeur moyenne idéale, recoupe exactement la théorie des erreurs propre à la physique de son époque. Les différences inter-individuelles ne seraient donc qu'un effet de halo au travers duquel il faut pouvoir distinguer la forme parfaite du modèle. Par conséquent, on ne s'étonnera pas que Quételet[7] dispensa une grande énergie à calculer des statistiques sur les mesures de toutes les parties du corps, et même de certains aspects « moraux », identifiant ainsi les « penchants » naturels de l'homme moyen. Il voulait ainsi se donner les moyens de dessiner le portrait de l'homme parfait voulu par Dieu. Remarquons la volonté inférentielle du travail de Quételet : si la moyenne *objective* des différentes mesures portant sur un même objet réel n'est rien d'autre qu'un ajustement à sa mesure réelle, troublée par des circonstances accidentelles, alors la moyenne de la

7. Quételet est aujourd'hui généralement ignoré ou oublié par les psychologues. Il fut pourtant très célèbre au siècle dernier et les sociologues le considèrent encore comme un des pères fondateurs de leur branche. En effet, le passage de l'homme moyen à l'homme social était par trop séduisant : après avoir lu Quételet, on pouvait concevoir « la société » comme une nouvelle entité dont l'existence pouvait être considérée comme indépendante de celle de ses constituants. Les travaux de Durkheim sur le suicide sont caractéristiques de cette vision déterministe des choses, derrière lesquelles on pouvait discerner des causes constantes et analyser leurs effets au niveau macroscopique. Grâce à Quételet, la magie statistique prenait corps : derrière la diversité infinie des individus, il devenait possible de parler d'entités singulières, existant à un niveau « supérieur », dont les relations devenaient plus simples à modéliser (*cf.* Desrosières).

distribution des mesures d'un caractère *a priori* abstrait, objective la tendance centrale en l'identifiant elle aussi à une entité réelle.

- *La « régression » et la corrélation*

Le modèle mathématique du phénomène que Galton appellera *reversion*, puis *regression toward mediocrity* avait déjà été étudié 80 ans auparavant par le mathématicien *Legendre* qui s'attaqua en 1805 à un problème posé par les astronomes : étant donné un certain nombre de couples (x,y) d'observations (entachées d'erreurs) que l'on suppose liés par une équation linéaire, quels sont les paramètres optimaux de l'équation ajustée, permettant d'associer à tout x_i une valeur $f(x_i)$, *aussi proche que possible* du y_i que l'on trouverait si on le mesurait sans erreur. Legendre imagina une méthode encore utilisée aujourd'hui dans la construction des *droites d'ajustement*, la *méthode des moindres carrés*.

Le terme de *régression* fut introduit beaucoup plus tard en biométrie à la suite d'un changement radical d'intérêt scientifique, véritable révolution qui donna naissance à la psychologie différentielle. Depuis 1830, et surtout avec Quételet, on pensait que les moyennes de la plupart des caractères mesurés sur une population restaient stables d'année en année : la taille des gens, les taux de suicides, etc. ne changeaient pas : l'homme moyen se perpétuait, immuable.

Darwin et les premiers évolutionnistes, et bien entendu les *eugénistes* anglais avaient, pour leur part, d'autres préoccupations : leur attention se portait au contraire sur les *extrêmes* des distributions, le concept d'*homme moyen* n'avait pas grand intérêt pour eux, car seuls retenaient leur attention les « génies » et les « tarés ». Le modèle normal de l'erreur (dont l'espérance était précisément l'homme moyen) changeait alors radicalement de sens : peu à peu on ne parlera plus d'erreur, mais plutôt de variation ou de *diversité*. En ne s'intéressant plus – à la manière des sociologues – à la tendance centrale mais aux extrêmes de la courbe, c'est-à-dire non plus à ce qui unit les individus*, mais à ce qui les sépare,* l'évolutionnisme darwinien et ses dérivés eugénistes furent à l'origine des premiers développements de la psychologie différentielle.

C'est le biométricien eugéniste Galton[8] qui introduisit les notions de médiane et de quartiles dans l'étude des caractéristiques des populations. Son objectif était de construire un espace commun dans lequel il pourrait représenter tous les cas étudiés, de manière à les comparer entre eux. L'idée de la *standardisation* était née : bien plus performant que la réduction à l'*homme moyen*, ce concept nouveau permettait une description précise de

8. Galton fut aussi un des premiers constructeurs de tests. Soucieux de comparer divers groupes humains entre eux, il construisit un « test des facultés humaines » qu'il administra à près de 9000 personnes. Cet instrument prête aujourd'hui à sourire car soucieux de faire plaisir à son cousin (C. Darwin), Galton se crut obligé de mesurer tous les aspects de la vie personnelle, psychique, physique et quotidienne des individus : il les interrogea sur la ferveur de leurs convictions religieuses, leur opinion vis-à-vis de l'école, leur aspect physique, leurs qualités morales, leurs aptitudes à vivre conjugalement et toutes sortes d'aspects qui nous paraissent un peu incongrus de nos jours.

groupes humains, ainsi que des comparaisons et des classements indépendants des circonstances variables de l'évaluation. L'objectif des eugénistes anglais de la fin du 19e siècle était l'amélioration de la société par la sélection biologique des individus les plus « méritants ». Cette idée, héritée directement de la théorie initiale de Darwin (cousin germain de Galton), allait se concrétiser au début du 20e siècle dans le courant scientifique appelé *biométrie*, auquel on doit la construction de la plupart des outils statistiques modernes.

Principal initiateur de ce courant (et donc premier fossoyeur de l'*homme moyen* immuable de Quételet), Galton cherchait, dès 1870, des lois génétiques permettant de prévoir les caractéristiques acquises de génération en génération. Il commença par étudier les caractéristiques des grains de pois et découvrit le phénomène qu'il appela tout d'abord *réversion*. Il remarqua que si l'on découpait la distribution d'une caractéristique des parents (poids du grain) en « tranches » (n-tiles) égales, et que l'on en calculait la moyenne, alors la moyenne de la même caractéristique mesurée chez les enfants correspondants n'était pas exactement la même. Les groupes de parents de taille élevée par rapport à la moyenne, donnaient des enfants dont la taille était également élevée par rapport à leur moyenne, *mais pas autant que celle des parents*.

Intrigué par ce phénomène incompréhensible, Galton récolta lors d'une exposition internationale sur la santé (1884) les mensurations de près de 9337 personnes, hommes et femmes, parents et enfants adultes. Il répéta ses observations et en conclut que l'hérédité avait une tendance naturelle à rapprocher les caractéristiques extrêmes de la moyenne, de génération en génération. Ainsi un enfant d'adultes très grands (Galton calcula un « parent moyen ») sera aussi très grand, mais *un peu moins*... (en moyenne). Cette loi (biologique) de *régression génétique vers la moyenne* fut publiée en 1885. Elle est aujourd'hui considérée comme fausse, car résultant d'une grave erreur de raisonnement. Galton fut lui-même troublé par le paradoxe suivant : si l'hérédité rapproche les extrêmes de la moyenne, comment expliquer que la variance des tailles des enfants est pratiquement la même que celle des parents ? Pour en trouver la réponse, il dut s'adresser à des mathématiciens qui lui fournirent l'explication : elle résidait dans la notion de *corrélation*.

L'erreur de Galton fut de croire que, par exemple, la frange des parents les plus grands devait nécessairement correspondre (avoir donné naissance) à la frange des enfants les plus grands. Si tel avait été le cas, le phénomène qu'il appela régression n'aurait pas été observable. En termes modernes, on dirait que la corrélation entre les variables taille des parents et taille des enfants aurait été parfaite, c'est-à-dire égale à 1. Cette remarque fut formulée par des mathématiciens qui connaissaient le concept de corrélation introduit par le physicien et astronome français Auguste *Bravais*, qui en exprima la formule en 1846, mais ne lui donna pas de nom particulier. C'est le mathématicien et collègue de Galton, K. Pearson, fondateur de la statistique moderne, qui définit exactement l'indice qui s'appellera désormais le *coefficient de corrélation Bravais-Pearson*.

Le paradoxe de Galton se trouvait alors expliqué : si la taille des enfants ne pouvait pas être prédite exactement sur la seule base de la taille des parents, l'égalité des variances des distributions des deux variables ne pouvait s'expliquer que par la variance d'une quantité aléatoire, dépendante de la force du lien existant entre elles, appelée *résidu*. Le *modèle de régression* n'est donc devenu opérationnel qu'au début de ce siècle, quand bien même Legendre en avait écrit l'équation dès 1805, comme nous l'avons vu plus haut. Le terme de régression a toutefois survécu malgré son inadéquation, personne n'ayant réussi à lui trouver une alternative unanimement acceptée.

C. La notion de test statistique

C.1. Exemple théorique et définitions

La notion de test d'hypothèse semble principalement due à Pearson et apparaît à la fin du 19e siècle. Un test statistique (d'hypothèse) consiste en une mise à l'épreuve d'une hypothèse dans le cadre des relations du couple population/échantillon : si les données d'un échantillon confirment – aux aléas du tirage près – un modèle théorique formalisé dans une hypothèse dite « nulle », énoncée toujours à propos d'une population, alors notre confiance dans cette hypothèse va s'accroître ; dans le cas contraire, on la rejettera avec un certain risque d'erreur librement choisi et consenti.

- *Voici un premier test « intuitif »*

Exemple : reprenons notre exemple de l'urne sans qu'il soit possible d'en inspecter le contenu, elle contient, nous dit-on, un certain nombre de boules. Sur la base de cette maigre information, on nous propose de décider si elle contient :

- H_0 : autant de boules blanches (B) que de boules noires (N) ; c'est l'hypothèse nulle d'équiprobabilité.
- H_1 : que des boules blanches (B) et aucune boule noire ; c'est l'hypothèse « alternative ».

Une première manière (rudimentaire) de résoudre cette énigme est de sortir des boules une à une de l'urne, et dès l'apparition d'une boule noire, nous aurons la *certitude* de la véracité de l'hypothèse nulle. Mais ce procédé n'est pas économique, car il est possible de tirer un grand nombre de boules blanches avant de tomber sur une noire.

Supposons maintenant que nous sommes dans une situation qui est généralement la règle dans la réalité : l'obtention d'éléments permettant la prise de décision coûte un certain prix ! Ainsi, *les grand échantillons nécessitent de plus grands investissements que les petit*s, surtout en sciences humaines où la prise d'information prend beaucoup de temps et n'est parfois pas très commode, ni toujours bien accueillie.

De manière analogue, dans notre exemple des boules, supposons que le tirage de chaque boule coûte un certain prix, et l'intérêt d'un *test* consiste donc dans l'économie de ses moyens : quel est le nombre de tirages minimum permettant de décider entre les deux hypothèses avec de bonnes chances de « tomber juste ?

Par le biais du test, nous allons donc renoncer à l'acquisition coûteuse d'une certitude, au profit de l'acquisition moins onéreuse d'une *conviction*, aussi solide que possible.

Combien nous faut-il alors tirer de boules de l'urne, au minimum, pour pouvoir choisir entre H_o et H_1 ? Pour répondre à cette question, nous allons choisir de *tester* H_o :

- *Si* H_o *est vraie*, quelle est la probabilité de tirer une boule blanche ?
  elle est évidemment de p(1B) = .50

- *Si* H_o *est vraie*, quelle est la probabilité de tirer deux boules blanches successives après remise ?
  elle est de p(2B) = p(1B) · p(1B) = .25 (car événements indépendants)

- *Toujours si* H_o *est vraie*,
  la probabilité de tirer 3 boules B successives est $(1/2)^3 = .125$
  et la probabilité de tirer 4 boules B successives est $(1/2)^4 = .0625$
  et la probabilité de tirer 5 boules B successives est $(1/2)^5 = .0312$

On voit que la probabilité de tirer plus de 5 boules blanches successives devient très faible si H_o est vraie, il faut donc choisir un *seuil* au-delà duquel il ne devient plus possible de croire en l'hypothèse nulle. On a encore trois chances sur cent de tirer 5 boules blanches successives si H_o est vraie, mais il ne reste que 1.5% de chances de tirer 6 boules blanches dans cette hypothèse.

Il semble raisonnable d'admettre (mais ce n'est qu'une convention) que si l'on tire 6 boules blanches successives de l'urne, celui-ci ne contient pas de boules noires. Cette décision est justifiée par le fait que l'événement : *tirer successivement 6 boules blanches d'un sac* est trop rare pour que l'on puisse croire qu'il contient une quantité égale de boules blanches et noires.

Cette décision est pourtant assortie d'un certain risque, car même si H_o était vraie, les probabilités de tirer 6 ou 7 ou 8 ou n boules blanches ne sont jamais vraiment nulles.

Toutefois, ce type de test « intuitif » ne correspond pas aux situations que l'on rencontre dans la réalité de la recherche où l'on est forcé de tirer en une seule fois un échantillon de taille définie. Les résultats fournis par l'étude de ce seul échantillon doivent alors servir de base pour la décision en faveur ou contre H_o, c'est la situation de test standard.

- *Test statistique standard (selon Fisher), test du modèle d'équiprobabilté ; par exemple H_0 : p(B) = p(N) ; pas d'hypothèse alternative.*

On nous propose à nouveau de décider si une urne contient des boules blanches et noires en quantités égales. Aucune autre hypothèse précise n'est énoncée, il faut simplement décider si H_0 est acceptable ou non en regard des données fournies par un « échantillon » tiré de l'urne. La taille de cet échantillon peut donner lieu à de longues discussions, mais admettons qu'il nous soit permis de tirer 10 boules de l'urne, en remettant chaque fois la boule tirée (tirage avec remise = condition d'indépendance des tirages).

Pratiquons l'*expérience aléatoire* et tirons 10 boules, observons les résultats : il y a 3 boules blanches dans notre échantillon, que penser alors de H_0 ?

On se rend bien compte que l'on pouvait trouver entre 0 et 10 boules blanches dans notre échantillon, avec davantage d'espoir d'en trouver 4, 5 ou 6 si H_0 était vraie.

En fait, ce qui nous manque, c'est la *distribution échantillonnale* de la variable : « nombre de boules blanches figurant dans un échantillon de 10 boules tirées d'une urne contenant autant de boules blanches que de boules noires ». Cette variable est aussi appelée *variable de décision*, puisque c'est sur la base de la valeur qu'elle prend lors de notre unique expérimentation que nous nous basons pour prendre une décision vis-à-vis de H_0.

Or il se trouve que les statisticiens ont trouvé une loi permettant de connaître la probabilité d'apparition de 0, 1, ... 10 boules blanches dans une situation telle que la nôtre. Il s'agit de la loi binômiale, qui donne les probabilités suivantes (pour une taille d'échantillon 10 et une proportion ω = 50%, *cf.* table A1.2 de Saporta) .

TABLEAU 1. : Répartition donnée par la loi binômiale pour n=10 et ω = 50%

k (nb. de B)	0	1	2	3	4	5	6	7	8	9	10
Prob.	0.001	.0097	.044	.1172	.2051	.246	.2051	.1172	.044	.0097	0.001
Prob. Cum.	0.001	.0107	.0547	.1719	.3770	.6230	.8281	.9453	.9893	.9990	1

Remarquons que cette loi donne aussi la répartition des occurrences de « Pile » ou « Face » lors de 10 lancers d'une pièce de monnaie. Notre problème de boules se réduit donc à celui qui consiste à savoir si une pièce est équilibrée (H_0) ou non.

Le principe du test statistique veut que si notre expérience (unique) fournit un événement « trop rare » sous H_0, alors nous aurons tendance à rejeter cette H_0 au profit d'une autre, encore non précisée. Qu'est-ce qu'alors un événement « rare » ? Par convention on admet que sont « *significatifs de la non validité de H_0 dans la population* » des réalisations de la variable de décision ayant moins de 5% de chances de se produire si H_0 est vraie.

Dans notre cas, les événements 0B, 1B, 9B et 10B sont très rares sous H_o. Plus précisément, la probabilité totale de l'un ou l'autre de ces événements est égale à 2 · .0107 = .0214 soit environ 2%.

Si l'on ajoute les événements 2B et 8B aux cas très rares, on obtient :
2 · .0547 = .1094 soit environ 11%, ce qui ne correspond plus au seuil fixé.

Ce qui signifie que nous pouvons observer entre 2 et 8 boules blanches dans notre échantillon de 10 sans pour autant devoir douter de H_o !

Par contre, si notre unique expérience fournit 0, 1, 9 ou 10 boules blanches, alors nous rejetterons H_o, au seuil α = 5% fixé par convention.

On notera que la décision de rejeter H_o si on trouve 0, 1, 9 ou 10 boules blanches est erronée 2 fois sur 100, puisque cette probabilité est précisément celle d'observer de tels événements sous H_o. (En principe, et dans tous les cas où la distribution de la variable de décision est continue, l'erreur de première espèce est égale à α).

- *Définitions*

- Les boules contenues dans l'urne mystérieuse constituent la **population** qui nous intéresse,
- H_o est une **hypothèse nulle** émise à propos de cette population,
- Les boules que nous pouvons tirer constituent un **échantillon**,
- L'acte de tirer cet échantillon est une **expérience aléatoire**,
- Le nombre de boules blanches observées à l'occasion de toutes les expériences aléatoires du type « tirer n boules » est une variable aléatoire appelée **variable de décision** ;
- Notre unique expérience, correspondant à *une* expérience aléatoire bien particulière, fournit un nombre (= le nombre de boules blanches dans les 10 tirées) qui est une **réalisation** de la variable de décision pour l'expérience en cours ;
- La répartition de la variable de décision est connue et tabulée, si bien qu'il est possible de définir un **seuil ou domaine de rejet** que la valeur de la réalisation de la variable de décision (dans notre expérience) ne doit pas dépasser, sous peine d'invalider H_o ;
- La probabilité cumulée d'observer des événements dépassant le seuil de rejet est égale au **niveau α de signification** du test (α = 5% en général) ;
- Par conséquent, α est aussi la probabilité de commettre une **erreur de première espèce**, *i. e.* rejeter H_o alors qu'elle est vraie ;

- La probabilité de l'**erreur de deuxième espèce** (β), *i.e.* ne pas rejeter H_0 alors qu'elle est fausse, n'est pas calculable dans ce cas et n'est pas un concept défini dans la perspective fishérienne ;

- La **puissance du test** ($1 - \beta$), *i.e.* la probabilité de rejeter H_0 alors qu'elle est effectivement fausse n'a pas non plus de sens dans ce contexte.

C.2. Signification de la signification statistique

Dans le jargon statistique, la *signification* doit être comprise comme « signe de… ». Par exemple : l'effet que j'observe dans un échantillon constitue *un signe* de l'existence de cet effet au niveau de l'intégralité de la population. Alors que dans le sens commun, la signification fait référence au sens, à l'intérêt ou à l'ampleur. *Une augmentation significative* est donc, pour le statisticien, une augmentation observée au niveau d'un échantillon, et suffisamment grande - relativement à sa taille – pour en inférer que cet effet peut être généralisé à toute la population.

Au sens commun, *une augmentation significative* est une forte augmentation… rien de plus. La confusion entre ces deux utilisations du même mot est regrettable car un effet peut être statistiquement significatif, tout en étant insignifiant. La signification statistique d'un résultat est donc une condition nécessaire mais non suffisante pour mériter d'être considéré avec attention :

- La condition est *nécessaire* parce que si le résultat n'était pas significatif, l'effet ou l'écart observé ne peut être attribué à autre chose qu'au hasard de l'échantillonnage, il est donc vain de l'interpréter.

- La condition est *non suffisante*, car les tests effectués sur de grands échantillons aboutissent pratiquement toujours à des résultats significatifs pour la simple raison qu'une hypothèse nulle correspond en fait et statistiquement parlant, à un événement *impossible*. En effet, par exemple l'événement : « trouver deux moyennes strictement égales dans deux échantillons » est un événement dont la probabilité d'occurrence est nulle…

- La signification statistique est donc surtout intéressante à considérer lorsque les échantillons sont petits, car dans ce cas, les aléas d'échantillonnage peuvent largement affecter la valeur des estimations. Il est important dans ce cas de savoir si on travaille sur un effet attribuable au hasard, ou non. Lorsque la taille des échantillons est respectable ($n > 100$), la notion de signification perd de son intérêt au profit de celle de taille de l'effet. Par exemple, si l'on étudie les liens existant entre des variables mesurées sur des échantillons de taille $n > 1000$, pratiquement toutes les corrélations calculées sont significatives, mais le véritable travail de recherche consiste à interpréter la différence entre une corrélation (significative) de .065 et une autre (également significative) de .84. Il est clair que la première

mesure nous incite à en déduire l'*inexistence* d'un lien (même si H_o a été vigoureusement rejetée !), et cela jusqu'à preuve du contraire.

Rappelons enfin (*cf.* Capel & al. 1996) a donné lieu à de vives polémiques :

- Défenseur d'une conception « fiduciaire » des probabilités, Fisher n'a jamais défini la signification vis-à-vis d'un seuil fixé d'avance, pour lui la p-value, ou probabilité d'occurrence d'un résultat de recherche sous l'Hypothèse nulle, est simplement une mesure du degré de fiabilité de celle-ci, *a posteriori*.

- Par contre, les mathématiciens Neyman & Pearson (le fils de Karl) ont défini le test statistique comme une véritable mécanique décisionnelle, dans le cadre d'une théorie fréquentiste de la probabilité. Pour eux, la variable de décision doit être clairement partitionnée en un domaine dit de l'« acceptabilité provisoire de H_o », et un autre dit « de rejet au profit d'une autre ». Dans cette conception, la notion de seuil prend tout son sens, ainsi que celui de probabilité d'erreur. Dans l'optique fréquentiste, celle-ci se définit simplement comme la probabilité cumulée d'observer des événements très improbables sous H_o.

Le point suivant explicitea plus clairement les problèmes posés par la mauvaise compréhension de ces conceptions originales, ainsi que les moyens de dépasser le nivau de la polémique en adoptant un point de vue raisonnable, loin des pratiques parfois presque superstitieuses ou « magiques » liés à l'utilisation systématique des tests d'hypothèse.

C.3 Du bon usage des tests d'hypothèse

Depuis leur popularisation par Fisher dans les années 30, les tests d'hypothèse ont été de plus en plus utilisés et constituent de nos jours un outil incontournable permettant la construction du savoir en sciences humaines, médecine, géographie et bien d'autres disciplines scientifiques. Cependant, quelques voix discordantes, en nombre croissant depuis les années 1980, ont mis en doute la bonne utilisation de ce type de technique. Ces dernières années, on a même vu certains auteurs demander que cesse l'usage déclaré « abusif » des tests d'hypothèse en sciences humaines notamment (voir à ce propos : Capel & al. 1996). Entre 1990 et 2000, la situation ne semble avoir guère évolué dans la pratique et en dépit de critiques en nombre croissant, aucun changement décisif ne se profile. Certains auteurs (Gigerenzer, référence *in* Capel, 1996) allant même jusqu'à considérer l'usage abusif des tests d'hypothèse comme une condition du développement d'un certain corpus de connaissances, particulièrement en sciences humaines, géographie et médecine (psychiatrie), pour ne citer que les domaines où les usages « pervertis » semblent être les plus répandus.

Un article récent (Tryon, 2001) montre qu'après le tournant du siècle, le problème n'a encore trouvé aucune solution. Dans un paragraphe intitulé « *the human factor problem* », l'auteur s'attache à décrire la situation critique qui est celle de la plupart des chercheurs

en sciences humaines. On peut en effet s'imaginer l'état d'esprit d'un chercheur débutant lisant ces lignes plutôt effrayantes, concernant le mauvais usage des NSHT[9] : « [...] *prominent investigators publishing in our best peer-reviewed journals for at least 3 decades have consistently misused NHST procedures* » ; ce qui signifie que : « [...] *editors and rewiewers who published these articles did not catch these mistakes* » ; et pire encore : « *NHST procedures are mistaught in at least six books written by leading psychometricians* » et pour déstabiliser définitivement tout nouveau chercheur en psychologie : « *Authors of nearly two dozen introductory psychology texts published between 1965 and 1994 err in their presentation of NHST procedures* ». Tryon remarque finalement que non seulement toutes les tentatives entreprises depuis quelques décennies pour corriger ces mauvaises conceptions des tests d'hypothèses se sont révélées vaines, mais que tout effort supplémentaire est sans doute également inutile[10](...).

Pour résumer ce très rapide survol de la question, nous noterons que depuis longtemps, quasiment depuis son introduction, l'outil « test de signification » appelé aussi « test d'hypothèse » est considéré, du point de vue de très nombreux auteurs, comme étant mal utilisé et mal compris, surtout par les personnes non formées en statistiques (psychologues, sociologues, géographes, psychiatres, etc.). Rappelons l'article de Hunter (1997) qui appelait à la cessation immédiate de l'usage de cet outil notoirement *perverti*. Dans un article essentiel, Gigerenzer (1993, cité in Capel & al. 1996) tentait d'analyser les causes « psychologiques » de ces problèmes d'interprétation, en même temps que les raisons du caractère incroyablement persistant de ces pratiques, mettant en cause d'une part l'implacable injonction « publish or perish » sévissant dans les milieux professionnels de chercheurs, et d'autre part les origines extrêmement conflictuelles qui ont présidé à la naissance des tests d'hypothèse, conflits qui sont généralement occultés (ou ignorés) par les enseignants de ces techniques. Le fait est que pour un chercheur en sciences humaines, il est bien souvent très difficile d'y voir clair, c'est-à-dire de savoir exactement ce qu'il « fait faux » lorsqu'il utilise des tests d'hypothèse, et pire encore ; peu lui est enseigné pour remédier à ces problèmes, excepté quelques citations de Cohen (1988) qui font allusion à une mystérieuse « analyse de puissance » dont tout le monde semble avoir entendu parler, mais que bien peu appliquent réellement.

Selon Gigerenzer, il s'est installé un climat de vague culpabilité qui est propice au dogmatisme scientifique. La chasse à la p-value « significative » constitue, ni plus ni moins, un impératif catégorique, *vital* : si $p < 0.05$ – *publish*; mais si $p > 0,05$ – *perish* ! Et l'on comprend alors mieux pourquoi il est si difficile – voire impossible – d'éradiquer les pratiques douteuses liées à l'usage des tests statistiques. Pour reprendre les termes acides de Salsburg (1985), elles se sont imposées comme une véritable religion, gage

9. NHST : Null Hypothesis Significance Testing.

10. Tryon utilise ce dernier argument pour introduire une nouvelle manière de contourner l'usage des NHST, technique qui ne nous intéressera pas ici.

d'emplois et de salaires pour d'innombrables chercheurs, universitaires, éditeurs, imprimeurs et autres professions associées, tous intéressés à ce que $p < 0.05$.

Les lignes qui suivent ont pour but de montrer que $p < 0.05$ n'est en fait pas une garantie de l'intérêt d'un résultat ou d'une recherche, mais que l'usage intelligent des tests d'hypothèse est possible (et pas si difficile) à condition d'apprendre à se servir d'un logiciel d'analyse de puissance (ou de posséder quelques notions de programmation). Bien plus, cet effort peut constituer la source d'un intérêt nouveau, loin du climat de culpabilité malsain, propre à l'usage mécanique et vide de sens des « valeurs p » et autres étoiles simples, doubles ou triples, directement inspirées d'un célèbre guide gastronomique.

En préambule, rappelons que les tests d'hypothèse actuellement utilisés sont une créature hybride (Gigerenzer, 1993) dont les auteurs « parents » sont mystérieusement occultés, et pour cause : aucun des deux n'y reconnaîtrait son petit. On attribue la paternité des tests de signification à l'agronome-mathématicien Fisher qui proposa cette technique (déjà utilisée par Pearson) pour se faire une idée intuitive de la crédibilité d'une hypothèse. Une hypothèse « nulle » concernant un modèle valable dans une population est mise à l'épreuve dans un test de signification effectué sur un échantillon représentatif (tiré en principe aléatoirement) de ladite population. La valeur observée, réalisation de la variable échantillonnale pour l'expérience en question, ne devrait pas, *si H_o est vraie*, s'éloigner « trop » d'une valeur attendue, donnée par une table. Cette technique suppose que la distribution de la variable échantillonnale est connue et tabulée (sous une forme standardisée), ce qui permet de connaître précisément la probabilité d'apparition (exprimée en centile) d'une valeur observée. Fisher (1935) déclarait volontiers qu'une valeur empririque dépassant le percentile 95 ($\alpha = 5\%$) de la distribution de la variable échantillonnale (loi normale réduite, chi carré, t, ou F) jetait le doute sur l'hypothèse nulle *et incitait à poursuivre l'expérimentation avec d'autres échantillons*. En utilisant cette technique intuitive, ou pouvait peu à peu affiner le modèle, de proche en proche, en adaptant les hypothèses et en répétant les expériences autant de fois que nécessaire.

Telle était la méthode inférentielle de Fisher, dans laquelle la valeur p (= 1 – percentile de la réalisation de la variable échantillonale pour l'expérience donnée) représente évidemment $p(D/H_0)$ c'est-à-dire la probabilité des données, *étant donné H_o*, et non pas le contraire $p(H_0/D)$, probabilité que l'hypothèse nulle soit vraie, étant donné les données, erreur fréquemment rencontrée, dont nous reparlerons. *Remarquons également que pour Fisher, il n'est question ni de décision*[11]*, ni de risque d'erreur, ni bien sûr de puissance d'un test*.

Pour les mathématiciens Neyman et E. Pearson (le fils de Carl Pearson associé au coefficient de corrélation avec son prédécesseur français Auguste Bravais), l'attitude

11. on ne rejette pas vraiment l'hypothèse : on en *doute* plus ou moins en mettant en évidence un *désaccord* entre les données (que Fisher appelle les *faits*) et celle-ci.

« fiduciaire » de Fisher ne pouvait pas donner lieu à une véritable construction de savoir scientifique. Pour ces esprits plus tranchés, une hypothèse ne peut pas être que « plus ou moins recevable », elle est doit nécessairement *être admise* comme vraie ou fausse : il s'agit donc de *décider*. Neyman et Pearson mirent au point en 1928 la forme achevée du « test d'hypothèse » dont certains éléments nous sont encore familiers de nos jours.

La conception fréquentiste des probabilités, partagée par Neyman et Pearson, donna naissance à la notion de risque quantifiable. Alors que Fisher déclarait : « *Nous nous trompons rarement en adoptant comme limite conventionnelle 0.05* [...] », la conception fréquentiste veut que la limite de 0.05 détermine précisément une zone « critique » incitant au rejet de l'hypothèse nulle, susceptible de conduire à exactement 5% d'erreurs de décision *sur le long cours*. De plus, le simple rejet d'une hypothèse ne conduisant pas à une conclusion satisfaisante et en aucun cas à la possibilté d'une décision, Neyman et Pearson introduisirent l'*hypothèse alternative*, forçant ainsi le chercheur à déterminer plus ou moins exactement à définir *l'écart qu'il s'attend à voir décelé par le test*. Le test d'hypothèse ainsi défini, on est en présence d'une véritable *mécanique décisionelle* dans laquelle les états d'âme du chercheur n'ont plus aucune place : celui-ci doit, *avant* de commencer l'expérience, définir un seuil de rejet α (définissant ainsi le risque de première espèce, à savoir la *probabilité de rejeter H_0 à tort*), décider d'un écart Δ intéressant pour sa discipline, écart pouvant être par exemple déterminé par deux moyennes alternatives caractéristiques de deux populations différentes ($\Delta = \mu_1 - \mu_2$). Cela étant fait, il doit encore décider de la *sensibilité* du test, c'est à dire adapter ses caractéristiques à la taille de l'écart (ou de l'*effet*) devant être décelé. Pour ce faire, il doit déterminer un seuil β, définissant le risque de seconde espèce qui représente la *probabilité de ne pas rejeter H_0 alors qu'elle est fausse*. Cette probabilité β permet immédiatement de connaître la puissance prévue du test, ou *probabilité de rejeter H_0 à bon escient* (égale à $1 - \beta$).

Finalement, l'équilibrage de tous ces paramètres exige également un ajustement de la taille de l'échantillon, car β est une fonction du degré de chevauchement des distibutions échantillonnales sous H_0 et H_1, le chevauchement étant déterminé par les écart-types de celles-ci, écart-types d'autant plus minces que la taille de l'échantillon est grande.

Voici en quelques mots une description sommaire, mais suffisamment fidèle, de la conception du test d'hypothèses selon Neyman et Pearson. On comprend aisément que les articulations logiques d'un tel raisonnement peuvent paraître lourdes et complexes, ce qui peut expliquer que bien peu de chercheurs appliquent cette méthode à l'heure actuelle sous cette appellation et c'est au contraire la conception hybride critiquée par Gigerenzer qui prévaut généralement.

Forts de la connaissance des origines, il devient dès lors plus facile de comprendre l'état actuel de la question. Selon Huberty (1993) et bien d'autres auteurs, les tests d'hypothèse utilisés depuis bientôt 40 ans ne sont en effet ni directement fishériens, ni réellement fidèles aux directives rigoureuses de Neyman et Pearson. Leur nature est en

vérité « hybride » car ils nient leurs origines tout en réalisant des confusions regrettables. Huberty remarquait en effet que très peu de manuels citent les « pères » (Fisher et Neyman + Pearson), comme si ces techniques existaient *sui generis*, incitant l'utilisateur à croire qu'elles seraient héritées, immuables, d'une tradition séculaire, gage de qualité et de sécurité absolue d'utilisation. La vérité est pourtant toute autre ; on sait que Fisher et Neyman & Pearson travaillaient dans le même institut et entretenaient des rapports conflictuels, au point que ces derniers furent obligés de poursuivre leurs recherches en Amérique. Les équipes des deux courants en conflit évitaient de boire le thé au même moment, etc... Les anecdotes piquantes ne manquent pas à ce sujet. On peut reconstruire sans trop de peine cette petite histoire (*cf.* Peters..) qui devrait nous rappeler que le traitement des tests d'hypothèse, et le traitement des méthodes inférentielles en général, *n'a jamais été l'objet d'un consensus* et a toujours été à l'origine de conflits intellectuels aigus. Cette réalité explique sans doute pourquoi les techniques inférentielles utilisées ces dernières années sont non seulement *hybrides*, dans une tentative déséspérée de concilier les pères ennemis, mais aussi *orphelines*, dans la mesure où l'impossibilité de concilier l'inconciliable faisait préférer l'oubli des origines à l'aveu de l'impossibilité de proposer une doctrine consensuelle. Cette situation pour le moins étrange dans l'histoire de la science ne pouvait que provoquer certains dérapages, précisément ceux qui sont décriés par toute une foule d'auteurs auxquels nous avons déjà largement fait allusion.

Voici donc en quoi consiste la pratique *hybride*, cible des critiques de Gigerenzer et de bien d'autres. Face à cette réalité, trois types de réactions sont possibles et peuvent s'observer en examinant les revues de littérature.

La première approche, de loin la plus courante, est celle des manuels de statistiques et d'analyse de données en sciences humaines, qui – et on devrait s'en étonner davantage - sont extraordinairement nombreux sur le marché, comme si chacun de leurs auteurs pensait qu'il est seul à vraiment être capable d'expliquer des techniques que pratiquement tout le monde utilise et croit connaître. Ces approches que l'on pourrait qualifier de « pédagogiques » tendent à rapprocher le sens des tests d'hypothèse des conceptions de Neyman et Pearson. Cependant, dans la mesure où les hypothèses alternatives restent vagues (du genre : H_0 : r = 0 et H_1 : r¦ différent de 0), l'attitude fiduciaire de Fisher reste à l'honneur, H_1 n'a en effet pas d'intérêt en soi et ne représente rien d'autre que la négation de H_0. Ces conceptions, qui font intervenir des hypothèses alternatives non spécifiées, ne sont pas très éloignées des conceptions *hybrides* évoquées ci-dessus. En effet, dans la mesure où elles empruntent à la rigueur de la mécanique décisionnelle de Neyman et Pearson certains éléments typiquement fréquentistes (seuil, risque d'erreur) pour les mêler à l'idée fishérienne du test de signification, tout se passe comme s'il s'agissait de *forcer Fisher à prendre une décision face à ses données*, attitude qu'il a toujours refusé d'adopter. Cette volonté de retrouver une certaine rigueur en fixant des valeurs de probabilité liées à des *risques* ne devrait pas occulter le fait que les sciences humaines ont une conception beaucoup plus fishérienne que « neyman-pearsonienne » des probabilités. Mais en affirmant cela, nous sommes encore loin d'avoir en main la clef

de la compréhension de l'existence et de la ténacité des pratiques « illicites », car la vraie conception hybride va encore y mêler des conceptions baysiennes, très intimement ancrées dans les réflexes épistémologiques des penseurs de sciences humaines. C'est ainsi qu'on en est venu, notamment, à *probabiliser l'hypothèse nulle* sur la base des faits, suprême perversion, tant pour Fisher que pour les inventeurs du tests d'hypothèse. Expliquons-nous : probabiliser l'hypothèse nulle est une des erreurs les plus graves commentées par Schmidt & Hunter. Il s'agit de l'erreur qui consiste à croire que $p = p(H_o/D)$, probabilité de la vérité du modèle (incarné par l'hypothèse nulle), *étant donné les données* (alors qu'on se souvient que p n'est rien d'autre que la probabilité des données, étant donné l'hypothèse). Cette tendance irrépressible à vouloir probabiliser la véracité d'un modèle hypothétique explique bien le *culte de la p value* ridiculisé par Gigerenzer ; « *if p <.05, publish, if not, perish* ». Lié à ce culte de la *p value*, il existe un véritable rituel des petites étoiles (*, **, ***) accompagnant pratiquement tous les résultats statistiques, t de Student, corrélations, etc. Officiellement, ces étoiles sont censées informer le lecteur de la valeur p du résultat : une étoile indique que la probabilité du résultat (sous H_o) est inférieure à 0.05, deux étoiles qu'il est inférieur à 0,01 et trois étoiles apportent un luxe supplémentaire, le fin du fin. Que nous apprennent réellement ces étoiles ? La réponse est *rien*, sinon l'information que leur utilisateur ne sait peut-être pas que la p value est fonction de l'effectif de l'échantillon. Il est pourtant clair qu'une corrélation calculée sur 1000 individus a toutes les chances d'être « trois étoiles », alors que *la même* corrélation calculée sur 100 individus ne sera gratifiée que d'une seule étoile, et toujours la même calculée sur 30 personnes sera reléguée avec mépris à l'infâme condition de « *non significant* ».

Cela dit, il est certes parfaitement vrai qu'une corrélation estimée à partir de 1000 individus est bien plus stable (donc fiable) que la même calculée sur 100, mais la p value n'a rien à faire dans cette affaire : *la seule* information pertinente qu'elle nous apporte est que *toute* corrélation, si elle est non significative, ne peut *pas* s'interpréter comme l'indice, le « signe », d'un lien entre deux variables dans une population parente. On comprend donc mieux le sens des petites étoiles : il y en a d'autant plus que l'effet est grand, certes, mais leur nombre augmente également si l'échantillon est grand ! Donc, en présence d'une série de résultats calculés sur des échantillons de tailles différentes, on ne sait pas ce que signifient réellement les étoiles (grand effet ou grand nombre ?). Et dans le cas de plusieurs résultats calculés avec le même échantillon, par exemple une matrice de corrélations, le nombre d'étoiles n'indique rien d'autre que la taille des effets, *ce que l'on peut voir de manière bien plus précise en regardant directement les effets* (corrélations, t, F, etc.). Il nous semble donc inévitable de supposer que le chercheur qui affuble les éléments de sa matrice de corrélations de petites étoiles se livre à un rituel vide de sens dont il serait bien en peine d'expliquer le sens et la raison, mais qu'il juge incontournable.

En guise d'exemple, reprenons le tableau que nous avions commenté dans un article précédent (Capel & al., 1996). Il est certes un peu forcé, mais explique bien l'enjeu lié aux petites étoiles :

TABLEAU 2. Corrélations entre deux jeux de variables

	X_1	X_2	X_3
Y_1	–.52 ****	–.37*	.90*****
Y_2	–.48**	.50***	.02
Y_3	.82*****	–.29	–.29

Note :

*p < .02, df = 40.

**p < .001, df = 40.

***p < .0007, df = 40.

****p < .0003, df = 40.

*****p < .0001, df = 40.

N'est-il pas évident ici que le nombre d'étoiles n'indique rien d'autre que la hiérarchie des tailles de corrélations ? Comment se fait-il que des chercheurs scientifiques écrivant pour des pairs (l'exemple est réellement tiré d'une revue « scientifique ») puissent parler un langage aussi vide de sens, pour nous faire voir de manière indirecte des choses que tout le monde peut voir directement, et de manière beaucoup plus informative ? Fisher serait sans doute choqué de découvrir un tel forfait contre le bon sens, lui qui écrivait (1935) : « Pourvu que l'écart soit nettement significatif, *il est sans importance pratique que p soit .01 ou .000001* [...] »). Quand à Neyman et Pearson ils seraient sans doute très étonnés d'apprendre que de telles pratiques portent un nom qu'ils ont donné à une technique décisionnelle qu'ils ont voulu rigoureuse, pour en finir avec les attitudes fiduciaires de Fisher et substituer des *calculs de risques* au sentiment d'incertitude.

Une deuxième réaction face à l'évidence d'une généralisation de pratiques hybrides mal comprises est de le rejeter et par suite de préconiser d'autres manières d'exprimer des différences. Divers auteurs préconisent depuis quelques années de ne plus utiliser les tests d'hypothèse et de les remplacer par des calculs d'intervalles de confiance qui évitent d'avoir à calculer des p-values problématiques. Cette attitude a toutefois peu de chances de s'imposer et il semble qu'un nouveau type de compromis s'impose peu à peu : on a pu remarquer que les dernières versions des logiciels statistiques les plus courants affichent maintenant les tailles d'effet et les puissances *post hoc* (par exemple SPSS). Reconnaissons qu'il s'agit là d'un progrès notoire en ce qui concerne les tailles d'effet (même si l'on peut facilement l'évaluer sans trop de peine en convertissant t, F ou chi carré en un équivalent de coefficient de corrélation), mais il faut aussi admettre que le calcul de la puissance *post hoc* n'est pas très informatif, celle-ci sera en effet insuffisante

si le test n'est pas significatif, et suffisante si le test l'est ! Ce phénomène découle nécessairement du mode de calcul qui identifie la différence observée à la taille d'effet qui aurait dû être déterminée à l'avance.

Une troisième attitude consiste à reprendre la réflexion probabiliste à la base et de s'intéresser aux travaux de Cohen (1994), auteur d'un type d'approche souple et originale, pas trop complexe du point de vue mathématique, à savoir « l'analyse de puissance *a priori* » (*power analysis*). On pourrait dire, en simplifiant un peu, que cette théorie constitue une tentative de revenir aux conceptions originales de Neyman et Pearson, mais sans leur adjoindre des notions fishériennes : l'idée n'étant plus de savoir si un résultat est « significatif » ou non, mais de mettre en place un *détecteur de différences* calibré *sur mesure* pour mettre en évidence une différence à laquelle on a réfléchi préalablement, et qui nous intéresse. Cette approche implique l'usage de la notion de *puissance d'un test* qui est l'objet du point suivant.

C.4 L'analyse de puissance selon Cohen (1988 et ouvrages suivants)

Avant de définir la notion de puissance d'un test, traduisons - dans la mesure du possible - le jargon inférentiel de Neyman et Pearson en une langue plus accessible permettant de nous introduire à celle de l'analyse de puissance. Nous pouvons comparer un test d'hypothèse (et son expérience aléatoire associée) à un *tribunal* dont la fonction est de juger si un individu est innocent (H_o) ou coupable (H_1) en regard de certaines pièces et témoignages (les « faits » de Fisher). Il est clair que l'accumulation des faits incite à douter de l'innocence qui joue ici le rôle de l'hypothèse nulle : en droit anglais, la personne est *a priori* considérée comme innocente et il en va de même avec l'hypothèse d'une liaison (entre deux variables), elle est a priori supposée absente ! Au delà d'une certaine quantité (seuil critique α) de faits, la présomption d'innocence n'est plus tenable, *mais l'accusé n'avoue jamais et on ne trouve hélas jamais de preuves absolues* ! Nous ne sommes donc *jamais* certains de sa culpabilité, si bien que toute condamnation s'accompagne toujours du risque d'avoir condamné un innocent. Ce risque gravissime (être un tribunal injuste) est socialement mal toléré, donc minimisé (conventionnellement, α est fixé à 5 ou 1%). Cependant, cette précaution n'est pas suffisante pour garantir l'exercice efficace de la justice, car un *bon* tribunal ne doit pas seulement se prémunir contre le risque de première espèce, sous prétexte qu'il craint de condamner des innocents, *il doit aussi se donner les moyens de ne pas relaxer des crapules*, car cela l'exposerait à être un tribunal inutile ! Le risque de « passer à côté » d'un coupable (risque de seconde espèce : β) est, pour sa part, évalué en général à 10 ou 20% dans les manuels de statistiques. Cette valeur conventionnelle montre que les statistiques inférentielles semblent obéir aux impératifs sociaux et éthiques de l'Angleterre démocratique du début du XXe siècle et tout porte à croire que pour un tribunal, il est moins grave d'être inutile qu'injuste... Nous ne pouvons qu'approuver cette résolution, mais nous ne pouvons nous empêcher d'être surpris qu'il en aille de même avec les tests d'hypothèses.

La *puissance* d'un test est un nombre égal à la probabilité, lors d'une décision basée sur une expérience aléatoire, de rejeter avec raison l'hypothèse nulle ou - en termes juridiques - de condamner un coupable à bon escient. La puissance d'un test peut donc être associée à la capacité de celui-ci à détecter une différence existant entre deux modèles mis en compétition dans un test de type Neyman & Pearson.

Précisons ces termes : le test Fishérien classique, avec définition d'une H_o, mais sans hypothèse alternative, ne permet de juger que le risque de 1ère espèce, à savoir rejeter à tort l'hypothèse nulle et il est impossible d'évaluer la sensibilité du test, c'est-à-dire la probabilité de ne pas rejeter H_o à tort (erreur de 2ème espèce β) ou, ce qui revient au même, la probabilité de rejeter H_o avec raison.

En utilisant l'analogie avec un tribunal, le test Fishérien s'assure bien qu'un condamné innocent ne soit pas condamné à tort, mais ne se préoccupe pas de savoir si le tribunal se donne les moyens suffisants pour condamner effectivement un coupable. Or, un tribunal qui se préoccupe de minimiser l'erreur de 1ère espèce évite, certes, d'être injuste mais un tribunal qui ne se préoccupe pas de minimiser l'erreur de 2ème espèce risque bien d'être inutile, et on sait par expérience qu'un bon tribunal doit se prémunir de manière équilibrée aussi bien contre l'injustice que l'inutilité.

Le fait que la plupart des chercheurs en sciences humaines pratiquent les tests statistiques sans se préoccuper de β (ou de leur puissance = $1 - \beta$) a suscité ces dernières années une vive réaction de la part de théoriciens de logique statistique : selon eux, une pratique des tests d'hypothèse qui néglige les considérations sur leur *puissance* équivaut à pratiquer un rituel vide de sens : autant alors renoncer totalement aux tests puisqu'on ne se préoccupe pas de savoir s'ils sont utiles ou non.

Parmi les auteurs qui ont tenté de redonner aux tests statistiques leur « dignité », Cohen est le plus cité. C'est lui, en effet, qui est à l'origine de l'analyse de puissance (*power analysis*) qui peut, d'une certaine manière, être considérée comme une remise au goût du jour de l'approche de Neyman & Pearson.

Selon cette approche, un test statistique ne peut mettre en jeu qu'une seule hypothèse, il doit obligatoirement mettre en jeu *deux* hypothèses concurrentes précises. Il n'est donc pas question, comme on le voit souvent, de définir des hypothèses alternatives vagues du type : r = 0 contre r différent de 0. La puissance d'un test ne peut en effet être définie que si l'on dispose de deux hypothèses alternatives *précises* et elle est d'autant plus grande que β est petit, et il y a donc 3 manières de la contrôler :

- Plus α est grand, plus β est petit et donc grande est la puissance, mais on n'accepte quasiment jamais (?) que $\alpha > 0,05$ car le risque de rejeter H_o à tort est très mal vécu, un tribunal qui se respecte refuse avant tout d'être injuste…

- Comme on ne peut pas agir sur α, on peut faire varier le décalage entre les deux distributions correspondant aux deux hypothèses en concurrence : cette distance est

appelée *taille de l'effet* (*effect size*). On voit que si l'on demande à un test de détecter une différence importante, il sera plus sensible (puissant) que si on lui demande de détecter une petite différence (ce qui paraît normal…).

- On peut aussi jouer sur la taille de l'échantillon, car plus l'écart-type des distributions échantillonales est petit, moins les distributions se recouvrent, et plus β se réduit pour une taille d'effet égale.

Les chercheurs qui ne se préoccupent pas de la puissance de leurs tests et qui veulent absolument trouver des résultats significatifs jouent sur ce dernier phénomène : ils augmentent la taille de leurs échantillons (ou cherchent à disposer des groupes les plus grands possible) jusqu'à ce que leurs résultats deviennent significatifs. Mais en négligeant de réfléchir à propos de la puissance, ils ont aussi négligé de réfléchir à une taille d'effet pertinente et réellement intéressante : ils en viennent, en procédant de la sorte, à mettre en évidence des différences entre valeurs théoriques et échantillonales si petites, que les tailles d'effet au niveau des populations n'ont peut-être aucun intérêt !

Les conclusions à tirer de ce qui précède sont les suivantes :

- On ne devrait pas procéder à des tests d'hypothèse sans réfléchir auparavant à l'ampleur des effets attendus, autrement dit sans être capable de définir assez précisément Ho et son alternative H1. Nous verrons plus loin que ce n'est pas si difficile.

- Si la prise d'information est facile et ne coûte rien, les tests pratiqués avec de grands échantillons sont toujours plus puissants que ceux pratiqués avec des petits. Le tout est, lorsqu'on a de très grands échantillons, de savoir si les effets mis en évidence ont vraiment un intérêt pratique. Par exemple, calculer une corrélation significative de 0.12 sur un échantillon de 500 sujets n'a sans doute pas grand intérêt s'il s'agit de la corrélation entre deux tests.

- Si la prise d'information coûte cher, et parfois il peut être très coûteux d'ajouter ne serait-ce que quelques individus à un échantillon, et si de petits effets observés peuvent déjà être considérés comme intéressants et il est souvent très utile de pouvoir calculer *a priori* le nombre minimum d'individus pour disposer d'un test suffisamment puissant, capable de détecter une différence à laquelle nous pouvons donner sens dans une perspective théorique ou pratique.

- S'il n'est plus possible d'augmenter la taille de l'échantillon et que les données sont prises bien avant la phase de traitement, il peut être intéressant de connaître *post hoc* la puissance des tests qu'il est possible de pratiquer, étant donné une taille d'effet définie et une taille d'échantillon invariable.

- Un dernier cas de figure peut se présenter : étant donné une taille d'échantillon non variable et une puissance exigée à priori, par exemple $1 - \beta = .80$, on peut se

demander quelle taille d'effet minimum un tel test peut déceler avec une probabilité de .80, par exemple.

En résumé, on se rend donc compte que les grandeurs suivantes sont liées :

- le niveau de signification α du test,
- la puissance du test,
- la taille de l'effet à déceler et
- la taille de l'échantillon ;

ce qui signifie que si l'on en contrôle une, on fera nécessairement varier les autres. Le but de ces ajustements est de mettre en place un test suffisamment sensible pour mettre en évidence un effet déclaré pertinent et intéressant dans le domaine de la recherche. Lors de ces ajustements, certaines limites sont toutefois acceptées de manière plus ou moins conventionnelle :

- α est rarement inférieur à 5%. Il est en effet très mal perçu qu'un test se donne trop de latitude envers le risque de 1ère espèce : un tribunal peut difficilement se permettre d'être injuste !
- Dans les recherches qui se préoccupent du risque de 2ème espèce, on constate que β = .20 est assez couramment accepté, ce qui montre que ce risque (être un tribunal inutile) est quatre fois mieux toléré que le risque de 1ère espèce ! Il nous semble que l'on devrait en toute bonne foi se demander si ces conventions ne découlent pas directement du code éthique de la société britannique du début du XIXe siècle ?
- Ayant à l'esprit les deux contraintes précédentes, le plus simple (en théorie) est de jouer sur la taille de l'échantillon : ayant fixé une taille d'effet et des seuils a et b, il ne reste qu'à déterminer combien de sujets l'expérience doit comporter pour satisfaire aux exigences du chercheur. Procéder de la sorte consiste à effectuer une analyse de puissance *a priori*, qui est le moyen le plus économique et le plus efficace permettant de déclarer significatif, ou non, un effet préalablement bien défini. Pour illustrer cette manière de faire de manière intuitive, il suffit d'admettre que si l'on veut visualiser un objet céleste (la lune, pluton, un cratère lunaire, une constellation), il semble assez évident que l'on ne se servira pas des mêmes instruments d'observation selon les objectifs du chercheur.

Reprenons l'exemple du point A.2. : un chercheur est désigné pour tester l'efficacité d'une nouvelle méthode d'enseignement sur la moyenne générale à une branche scolaire, peu importe laquelle. Ayant préalablement accepté le risque de première espèce α de 5% (risque de déclarer «utile» la nouvelle méthode d'enseignement), mais aussi le risque β (conventionnellement fixé à 20% = risque de ne pas mettre en évidence les bénéfices de cet enseignement), et ayant décidé qu'il fallait déceler une différence d'au moins un demi point en moyenne sur une échelle de 6, entre le groupe «traitement» et le groupe

«contrôle», il ne reste qu'à calculer le nombre de sujets nécessaires à cette expérimentation. Cette manière de faire, l'analyse de puissance *a priori,* est préconisée par Cohen et c'est celle que nous conseillons vivement aux chercheurs.

C.5. Le problème de l'évaluation *a priori* de la taille d'un effet

Nous n'allons pas exposer ici les détails techniques nécessaires à la mise en oeuvre d'une analyse de puissance, le manuel de Howell (1998) est bien assez clair à ce sujet, mais il nous a semblé utile de préciser quelques concepts centraux de cette théorie, et les questions auxquelles ils sont liés.

La puissance du test est également liée à la taille de l'effet supposé être mis en évidence et le plus naturel semble que le cherche soit capable de fixer avant l'expérience une taille d'effet intéressante pour son domaine. (Pour un astronome, cette exigence revient à se demander quel type de lunette il va commander pour pouvoir correctement examiner l'ensemble du disque lunaire, par exemple ; et il va de soi que s'il veut observer une exoplanète située dans une constellation lointaine, son instrument lui coûtera beaucoup plus cher !). En sciences humaines la réponse à la question : « quelle taille d'effet est-il intéressant de mettre en évidence ?» n'est pas toujours simple. Dans le but d'aider les chercheurs, Cohen s'est attaché à clarifier au mieux les liens unissant α, $1 - \beta$, N (effectif de l'échantillon) et d (la taille de l'effet). Ces relations sont particulièrement faciles à comprendre dans le cas des tests portant sur des différences de moyenne. Dans un tel cas, Cohen définit d'abord la « taille de l'effet » comme la différence entre les moyennes théoriques attendues sous H_o et sous H_1, rapportée à l'écart-type de la population « parente ». Cette valeur nécessite la connaissance du sigma de la population qui est en principe théorique et donc inconnue, mais elle peut être facilement estimée en prenant le sigma de l'échantillon. Ainsi définie, d est indépendante de N, mais comme nous venons de le souligner ci-dessus, H_1 n'est pas toujours facile à définir, il faut alors estimer « intuitivement » d, opération que nombre de chercheurs répugnent à effectuer car ils estiment en général ne rien savoir de H_1.

Cohen prétend que tout chercheur peut se faire une idée, même imprécise, de la taille de l'effet attendu et il va même jusqu'à proposer 3 catégories d'effets :

- $D < .20$: petits effets, les distributions échantillonales sous H_o et H_1 se chevauchent à 85%,
- $.20 < D < .50$: effets moyens, 66% de chevauchement,
- $D > .80$: effets importants, 53% de chevauchement.

La taille de l'effet peut donc être soit calculée, si H_1 est bien précisée et si le sigma de la population est bien estimé par celui de l'échantillon, soit choisie au moyen des repères fournis par Cohen.

Le raisonnement et les calculs permettant d'expliciter les liens entre a, S, N et 1 - b ne sont pas simples, aussi est-il plus commode d'utiliser soit une table, soit un petit logiciel qui permettent de connaître la valeur de l'un des 4 paramètres en fonction des 3 autres (*cf.* table « annexe de la puissance » de Howell).

En conclusion, si l'analyse de puissance est peu pratiquée, c'est peut-être parce qu'elle n'est pas simple à comprendre et nécessite une vision très claire de ce qu'est (et n'est pas) un test d'hypothèse. Il faut aussi reconnaître qu'elle est relativement difficile à appliquer dans les cas autres que les tests les plus simples, comme les comparaisons de moyennes, tables de contingences, etc. Cependant la tendance actuelle montre que ces préoccupations entrent peu à peu dans les habitudes intellectuelles des chercheurs en sciences humaines, d'autant plus qu'il existe maintenant des petits logiciels de calcul très simples permettant de trouver facilement les paramètres nécessaires. Il ne nous reste donc qu'à proposer quelques exemples et exercices pour convaincre le lecteur de l'intérêt de cette approche et de son accessibilité (voir en fin de volume, appendice).

C.6. Quelques exemples d'application de l'analyse de puissance

- *Exemple 1. (Ajustement à une moyenne théorique)*

On s'intéresse au score moyen à l'échelle « Tension » d'un test de personnalité passé par des personnes exerçant la profession d'enseignants. Pour simplifier, admettons que les scores sont déjà standardisés et s'expriment sous forme «G», c'est à dire avec une moyenne théorique de 50 et un écart-type théorique de 20.

- H_o : le score moyen des enseignants est M_o = 50 = moyenne de la population générale
- H_1 : le score moyen des enseignants est M_1 = 60

Cherchons, pour une puissance 1 - β = .80, la taille nécessaire d'un échantillon permettant de mettre cette différence en évidence. Autrement dit, quelle est la taille de l'échantillon permettant de mettre en évidence une différence de 10 (dans le sens d'une augmentation, le test sera unilatéral) entre la moyenne du groupe et la moyenne théorique 50, avec un risque de première espèce de 5% et la garantie que si la différence est significative, alors le test va la mettre en évidence 8 fois sur 10 ?

Solution : le plus simple est d'avoir recours à un programme de calcul comme par exemple *G power* (gratuit, téléchargeable sur internet). Choisir l'option « t test - one sample test » et entrer les paramètres décidés *a priori* (type of power analysis - *a priori, compute sample size given α, power and effect size*) : α est toujours égal à 0.05 *one tail*, la taille d'effet est 0.5 (d = 10/20) et la puissance 0.8. Cliquer sur le bouton « calculate » et le programme affiche immédiatement les résultats qui montrent que 27 personnes suffisent pour réaliser les objectifs assignés au test.

Variante 1 : si on dispose a priori de 16 personnes et que l'on désire mettre en évidence la même différence, quelle sera la puissance du test ?

Solution : activer l'option *post hoc* de *G power* (compute power given effect size, α and sample size). Entrer l'effectif 16, garder α= 0.05, one tail et d = 0.5, *calculate* et la réponse apparaît : .60. Un tel test, peu sensible, passera à côté d'une différence significative (erreur de 2ème espèce) 4 fois sur 10 (β = 1 - .60 = .40).

Variante 2 : on décide de se contenter de 16 personnes et on vise une taille d'effet de .75 (M_1 = 65), la puissance sera-t-elle suffisante ?

Solution : la réponse est oui, pour une telle taille d'effet, la puissance de ce test sera proche de .90.

Variante 3 : on veut être pratiquement certain que notre test détectera une différence faible, par exemple M_1 = 54 (d = 0.20), combien de sujets seront nécessaires ?

Solution : on veut donc un test très sensible, par exemple de puissance 0.95. *G power* renvoie un effectif minimum de 272 personnes. Pour un test bilatéral il faut 327 personnes !

- *Exemple 2. (tests sur deux échantillons indépendants)*

Dans ce cas, le problème de la différence de taille des groupes n'est pas très important, *G power* gère très bien ces situations qui peuvent être source de complications si l'on se réfère à des tables. On s'intéressera par exemple aux *différences entre scores moyens* à l'échelle « Tension » d'un test de personnalité passé par des hommes et des femmes. Toujours pour simplifier, admettons que les scores sont déjà standardisés et s'expriment sous forme «G», c'est-à-dire avec une moyenne théorique de 50 et un écart-type théorique de 20 (population parente théorique).

- H_o : le sexe des participant/-es n'entraîne aucun effet ($M_F = M_H = \mu_o$) sur les scores de tension
- H_1 : le groupe d'hommes est en moyenne plus tendu que les femmes, la différence est de 5, soit un quart d'écart-type, donc d = 0.25.

Cherchons, pour une puissance usuelle de 1 - β = .80, la taille nécessaire d'un échantillon permettant de mettre cette différence en évidence. Autrement dit, quelle est la taille de l'échantillon permettant de mettre en évidence une différence de 5 (dans le sens d'une augmentation, le test sera unilatéral) entre la moyenne du groupe d'hommes et celui des femmes.

Solution : choisir l'option « t test - means - difference between two independant means (two groups) » et entrer les paramètres décidés *a priori* (type of power analysis - *a priori, compute sample size given* α*, power and effect size*) : α est toujours égal à 0.05

one tail, la taille d'effet est 0.25 (d = 5/20) et la puissance 0.8. À côté de « allocatio ratio » entrer le rapport entre les effectifs des groupes, a priori 1 si possible.
Cliquer sur le bouton « calculate » et le programme affiche immédiatement les résultats qui montrent que 51 personnes par groupes suffisent pour réaliser les objectifs assignés au test.
Si l'on ne pouvait disposer que de très peu d'hommes, par exemple 4 fois moins (allocation ratio = 4), le test renvoie les effectifs de 31 hommes et 125 femmes.

- *Exemple 3. (tests sur deux échantillons pairés - dépendants)*

Dans ce cas, le problème de la différence de taille des groupes ne se pose plus, on cherche à déceler une différence de moyenne entre deux passations d'un même test par un même groupe d'individus.

Solution : choisir l'option « t test - means - ... (matched pairs) » et procéder comme décrit ci-dessus.

En résumé, pour les t-tests il peut être intéressant de disposer d'un tableau récapitulatif donnant les effectifs requis pour diverses situations :

TABLEAU 3. Effectifs requis pour des tests de t sur un ou deux groupes, en fonction de trois types d'effets prédéfinis

Taille d'effet	*d*	*N = (1 éch.)*	*N = (2 éch.)*
petite	*0.2*	*196*	*784*
moyenne	*0.5*	*32*	*126*
grande	*0.8*	*13*	*49*

- *Le cas des tables de contingences*

G power permet aussi, en principe, de répondre aux questions de l'analyse de puissance pour les différences observées entre distributions de scores catégoriels (tables de contingences). Le problème, ici, est qu'il est bien plus difficile de se décider *a priori* au sujet de différences pertinentes ou non, surtout en sciences humaines. La situation est par exemple bien plus claire en botanique où les lois de Mendel permettent clairement de définir des modèles attendus, comme le montrera le chapitre suivant. Comme, de plus, *G power* n'est pas d'un usage très simple dans le cas des tables de contingences (il exige le calcul d'un paramètre de non centralité sans expliciter clairement le calcul), nous laissons provisoirement ce chapitre en suspens.

- *Différences entre deux proportions, deux corrélations, ou entre plusieurs moyennes :*

Le logiciel *G power* permet de répondre aux questions de l'analyse de puissance pour des différences entre proportions (z tests) et pour des ajustements de corrélations à la

valeur 0 ou toute autre valeur. Un grand nombre de variantes de l'analyse de variance (ANOVA et MANOVA) simple ou multiple est également pris en compte, mais l'étude complète de toutes les possibilités exigerait le volume d'un manuel spécifique sur ce thème.

C.7. Un exercice *décisif* sur l'analyse de puissance

Le lecteur désirant comprendre de manière approfondie le sens d'un test statistique et l'apport de l'analyse de puissance peut se livrer à l'exercice suivant :

- Générer une « population » U théorique de scores distribués normalement, de moyenne 0 et d'écart-type 1 (scores standards gaussiens), veiller à ce que N soit grand, disons 10000. Cette distribution est associée à l'hypothèse nulle H_0
- Générer une « population » X théorique de scores distribués normalement, de moyenne 0.5 et d'écart-type 1, N = 10000 aussi. Cette distribution est associée à l'hypothèse alternative H_1.
- Dans U, tirer aléatoirement 100 scores et pratiquer un test d'ajustement sur la moyenne théorique 0, au seuil α = 5% unilatéral. **Vérifier** que sur 100 tests de cette sorte, environ 5 donnent un résultat significatif (alors que H_0 est vraie puisqu'on a pris la population U parente). Noter que ce phénomène est indépendant de n !
- Ensuite : déterminer à l'aide *G power* la taille de l'échantillon nécessaire pour détecter une différence de 0.5 (taille d'effet correspondant à la différence de moyenne entre U et X), avec un seuil α de 5% et une puissance .80. On trouve n = 27. Dans X, tirer aléatoirement 27 scores et pratiquer un test d'ajustement sur la moyenne théorique 0, au seuil α = 5% unilatéral. **Vérifier** que sur 100 tests de cette sorte, environ **20** donnent un résultat **non** significatif (alors que H_1 est vraie puisqu'on a pris la population Z parente)
- On peut faire varier la puissance ou la différence entre les distributions U et Z et se rendre compte concrètement de ce que représente la *sensibilité* d'un test, quantifiée par la notion de puissance (plus un test est puissant, plus il est sensible).

D. Tests d'ajustement à des modèles théoriques

Les test d'ajustement ont pour but de comparer des résultats observés à ceux que l'on devrait obtenir théoriquement si un certain modèle *a priori* décrivait parfaitement une réalité donnée. L'objectif de toutes ces méthodes est donc de savoir (par le biais de l'information apportée par une expérience particulière) si un modèle théorique simple (une loi, une distribution de fréquences ou une valeur) peut décrire une réalité inobservable directement. Souvent, les modèles théoriques sont l'expression de certaines théories (comme par exemple la théorie chromosomique de l'hérédité) et le non rejet de l'hypothèse nulle peut-être interprété comme une confirmation empirique de celle-ci.

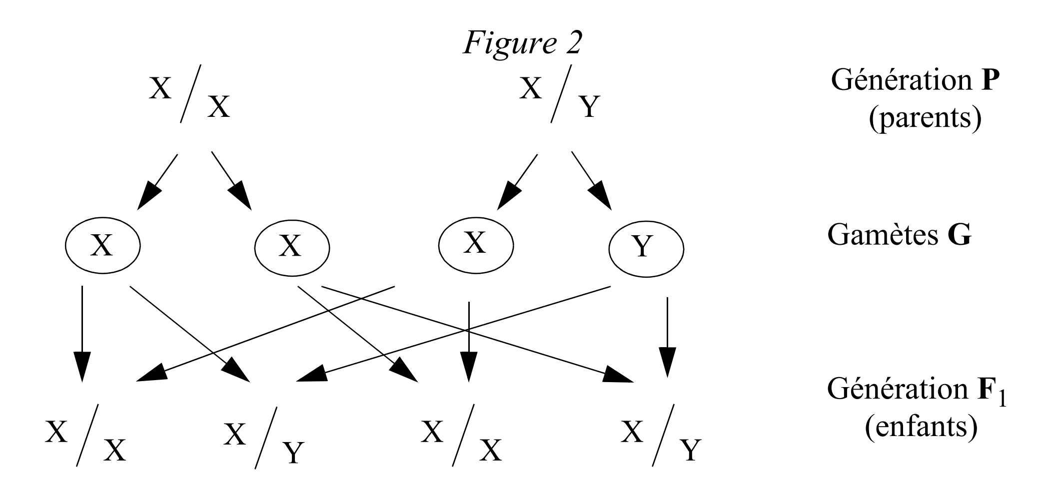
Contrairement à une croyance bien répandue, dans ce genre de test ce n'est pas le rejet de H_o qui est interprété comme un résultat intéressant, mais au contraire sa conservation.

D. 1. Introduction : le contexte de la naissance des tests d'ajustement

Vers 1900, De Vries redécouvre les travaux de Mendel (oubliés depuis 1865) concernant les lois[12] qui semblent régir la transmission des caractères simples au travers des générations. Reprenant les expériences classiques, De Vries leur appliquera une méthode statistique plus rigoureuse, celle des **tests d'ajustement** mis au point à la même époque dans les laboratoires anglais (notamment par K. Pearson). En voici un énoncé simplifié :

Les caractères héréditaires sont portés par les chromosomes, au nombre de 2n selon l'espèce animale ou végétale. Chez l'être humain, il y en a 46 (2 fois 23 de type X chez la femme, et 23X + 22X + 1Y chez l'homme. La transmission des caractères simples se fait par les chromosomes et obéit à un certain nombre de lois (dites de Mendel). En voici quelques illustrations :

Figure 2



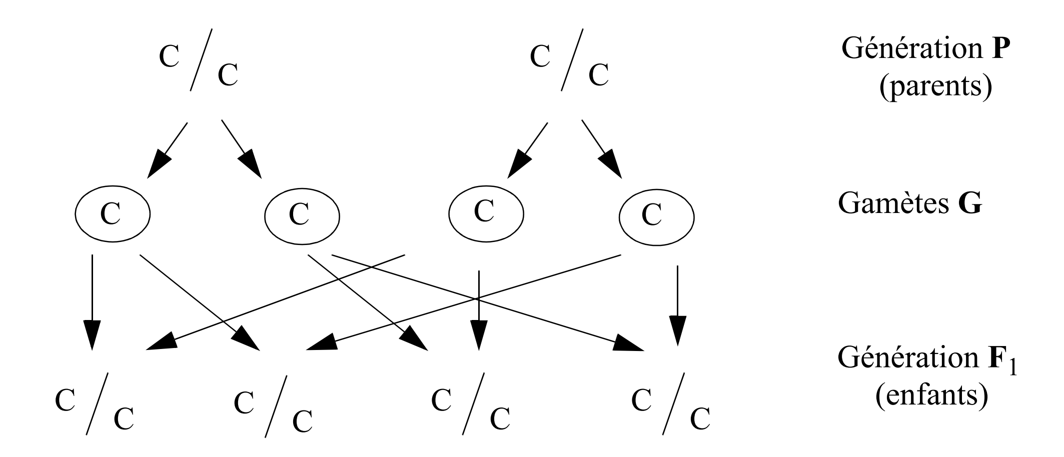
D.1.a. Transmission héréditaire du sexe :

On trouve bien (figure 2) la proportion généralement observée de 50% de garçons et de 50% de filles, mais cette proportion est théorique car dans un échantillon, celle-ci fluctuera autour de cette valeur attendue avec un écart-type plus ou moins grand. D'où l'intérêt de disposer d'une technique permettant de tester si la proportion observée ne s'écarte pas « significativement » de la valeur attendue (50%), auquel cas il faudrait remettre en question la théorie du mécanisme de transmission des caractères sexuels.

12. On notera qu'il a fallu attendre 1933 pour comprendre les mécanismes cellulaires mis en jeu (Roux et Weisman : théorie chromosomique de l'hérédité).

D.1.2. Transmission héréditaire d'un caractère C (indépendant du sexe) :

Figure 3



100% de sujets “purs” (homozygotes)

L'une des principales loi de Mendel est la *loi d'uniformité,* dans le cas de la transmission de deux caractères de « force » égales (cas des végétaux à fleurs, par exemple : C_1 = rouge et C_2 = blanc) on observe le phénomène suivant :

Figure 4

C_1/C_1 **rouge** C_2/C_2 **blanc** Génération **P** (parents)

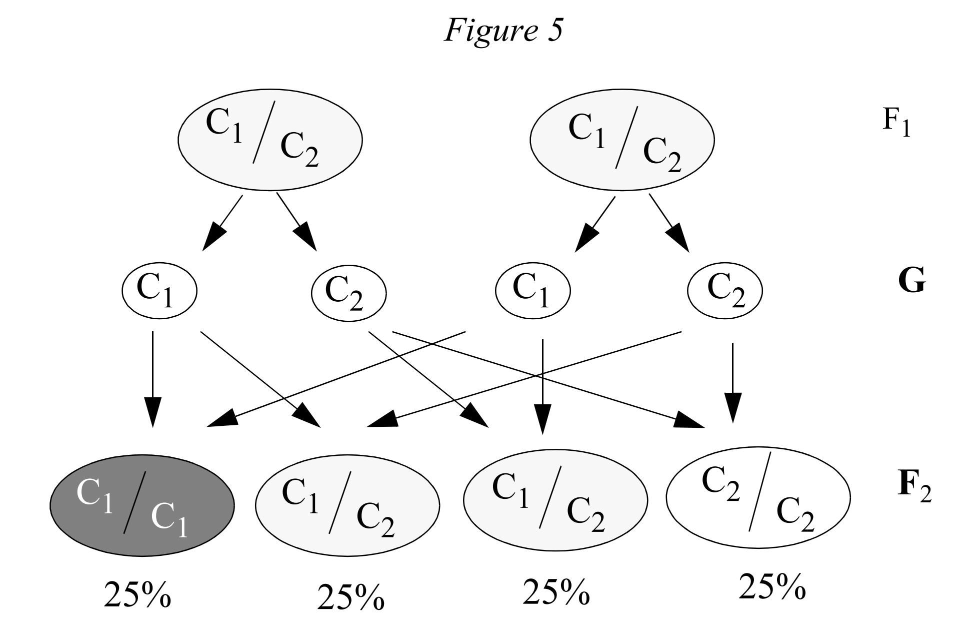
C_1 C_1 C_2 C_2 Gamètes **G**

C_1/C_2 C_1/C_2 C_1/C_2 C_1/C_2 Génération $\mathbf{F}_1$ (enfants)

100% de sujets « bâtards » **roses** (hétérozygotes)

La première génération est donc uniforme, mais si on continue à croiser les sujets de cette génération, on observe le phénomène suivant (*cf.* figure p. suivante) :

Figure 5



Les caractéristiques parentales réapparaissent à la seconde génération, dans une proportion de 1/4 rouge, 1/4 blanc, le reste étant rose. Ce phénomène ne s'explique que par la seconde loi de Mendel dite de la *pureté des gamètes* qui postule que les gamètes ne sont jamais hybrides (car haploïdes).

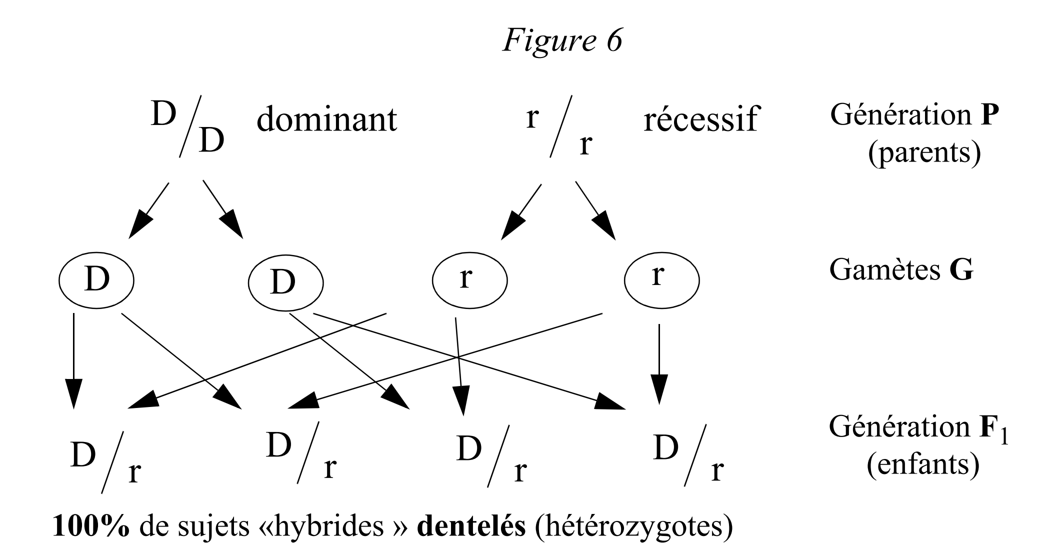
Une telle loi peut se vérifier par l'expérience. En croisant des fleurs rouges et blanches sur deux générations, on devrait, *si la théorie est vraie*, retrouver la *distribution théorique* des fréquences ci-dessus, soit dans ce cas : 25% de rouge, 50% de rose et 25% de blanc.

Au début de ce siècle, K. Pearson résolut mathématiquement ce genre de problème d'ajustement qui, dans des cas plus complexes requiert l'usage de distributions théoriques de type *chi-carré* qu'il fut le premier à calculer. On comprend du même coup pourquoi le développement de la statistique des tests d'ajustement statistiques est si étroitement liée au développement de la génétique (Pearson a été directeur pendant près de 30 ans du *Galton laboratory of genetics*).

D.1.3. Cas de dominance d'un caractère :

Dans ce cas (par exemple : D = feuille dentelée et r = feuille régulière), lorsque deux caractères différents sont présents dans le même génotype, c'est celui qui est exprimé par un gène (D/D) dit « dominant » qui occulte l'autre appelé dans ce cas « récessif ». Le caractère récessif (r) ne pourra donc s'exprimer que si le gêne dominant est absent, autrement dit si le génome du porteur comporte une paire de gênes récessifs (r/r). Comme le montre la figure suivante, les « enfants » de parents D/D (homozygote dominant) et r/r (homozygote récessif) sont tous hétérozygotes, de phénotype D :

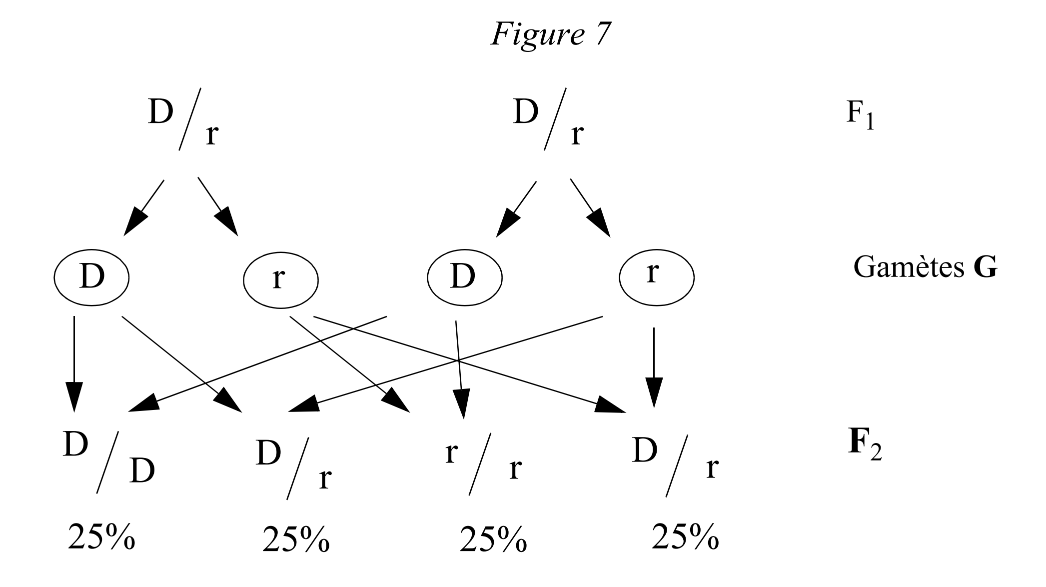
Figure 6



100% de sujets «hybrides » **dentelés** (hétérozygotes)

Si l'on croise maintenant deux hybrides D/r (*cf.* figure suivante), on observe 75% d'individus de phénotype « dentelé », dont 25% de génotype « dentelé pur » et 50% d'hybrides (hétérozygotes). On trouve par ailleurs 25% de feuilles « régulières pur ». Le gène récessif ne peut donc s'exprimer que dans 25% de la deuxième génération.

Figure 7



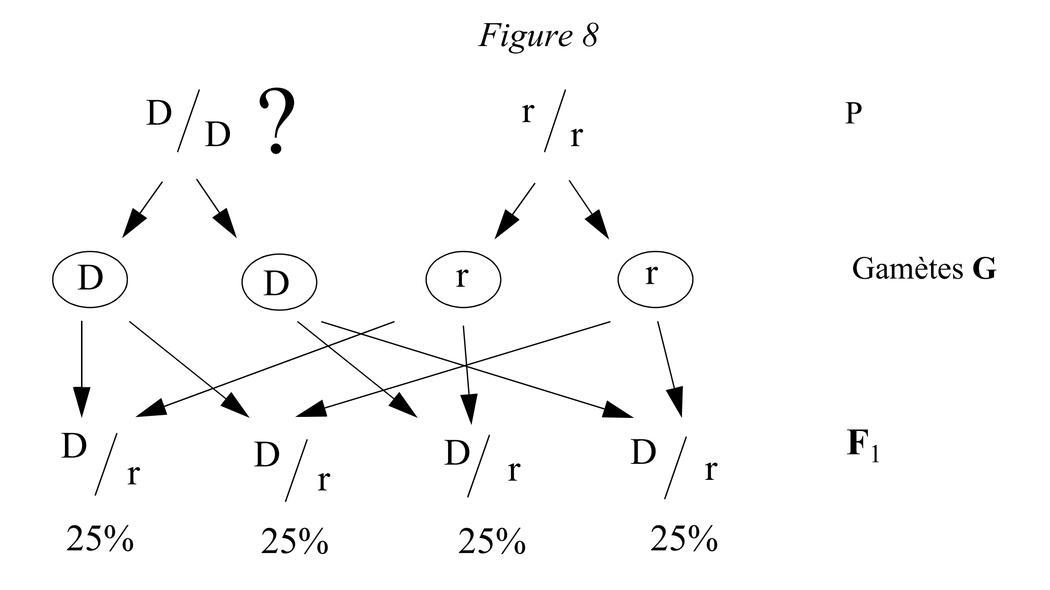
Problème important pour les généticiens du début du siècle :

Comment savoir, en présence d'un individu de phénotype donné, s'il est de race « pure » ou s'il est « hybride » ? Autrement dit : connaissant le phénotype (l'apparence), comment peut-on en déduire la connaissance du génotype (structure), dès lors que les génotypes D/D et D/r déterminent le même phénotype ?

Méthode astucieuse : croiser un individu dont on ne sait pas s'il est D/D ou D/r avec un homozygote récessif r/r.

- Cas 1. Si l'individu dont le génotype est inconnu est *homozygote* (génétique ment « pur ») on a :

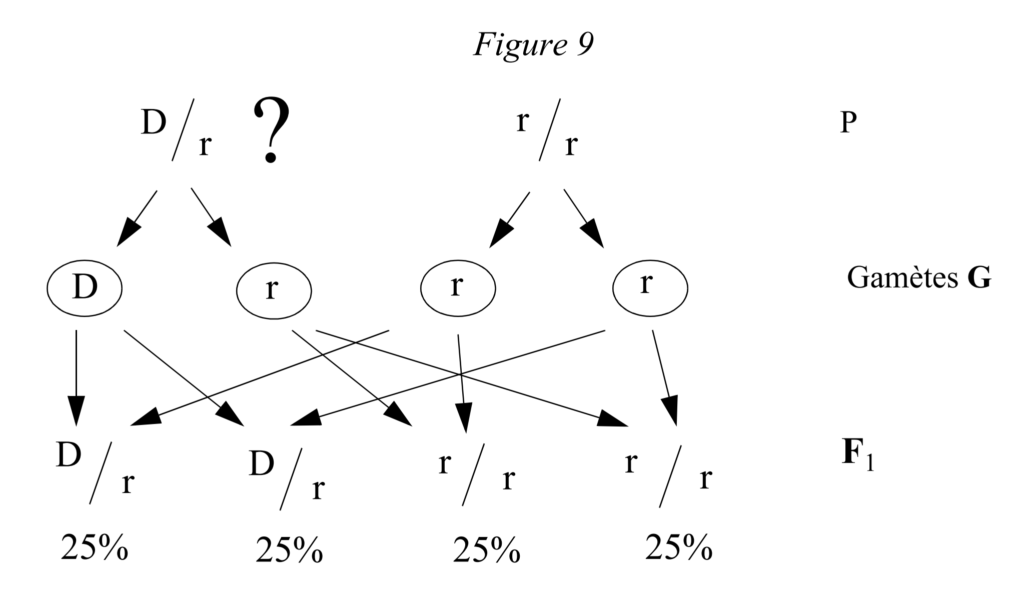
Figure 8



On ne retrouve plus de r/r, mais 100% d'hybrides dentelés !

- Cas 2 : Par contre, si le sujet inconnu est hybride :

Figure 9



On trouve 50% d'hybrides dentelés (D/r) et 50% de phénotypes récessifs « purs » de type r/r.

Du point de vue du plan expérimental, deux hypothèses sont en concurrence :

- (H_o) : la variété sélectionnée est de « race pure », et
- (H_1) : la variété sélectionnée est un hybride.
- À H_o correspond le modèle A : 100% de phénotypes D ; et

- à H_1 correspond le modèle B : 50% de D et 50% de r.

L'**expérience aléatoire** consiste à croiser des plantes dont le génotype est inconnu avec des plantes de génotype récessif r/r, puis de tirer au hasard des plantes dans la population F_1 de leurs « enfants ». On compte alors les occurrences du phénotype D. Si ce taux avoisine 100%, on acceptera l'hypothèse H_o, sinon on préférera H_1.

Ne serions nous pas ici en présence d'un « test d'eugénisme » ?

D. 2. Test d'ajustement à une distribution théorique continue, le modèle normal

Une bonne partie des tests d'aptitudes sont présentés dans les manuels accompagnés d'étalonnages normalisés, en particulier les tests de Q.I. Il n'est pourtant pas évident que certaines aptitudes soient automatiquement distribuées normalement dans toutes les populations, et il peut être intéressant de tester la normalité d'une distribution lorsqu'on dispose de données provenant de populations peu étudiées.

Les tests de normalité peuvent consister en diverses analyses de complexité variable.

- Le simple coup d'oeil distingue facilement des distributions très asymétriques, mais ne peut pas juger les écarts dus à la « voussure », ce genre de test empirique est donc insuffisant.

- L'analyse des paramètres de distribution est plus fiable, et on peut tester les coefficients de symétrie et de voussure (ou aplatissement), *cf.* Capel, Guide des T.P. p. 87.

- La plupart des logiciels d'analyse statistique effectuent sur demande un test de normalité avec analyse du Q-Q Plot et test du K-S de Lilliefors.

- Le principe des tests de normalité est basé sur la mesure de l'écart entre certains fractiles de la distribution observée, et les mêmes fractiles donnés par la loi normale. Ces différences sont mises au carré, rapportés aux valeurs attendues et sommées, et la valeur globale de la différence est une quantité D^2 qui suit une loi de chi carré à n-1 degrés de liberté, n étant le nombre de fractiles utilisés. (*cf.* Tables statistiques en annexe).

En résumé : test de normalité - mode d'emploi

- *Conditions d'utilisation* : si possible au moins 100 sujets tirés au hasard
- *Procédure à suivre* :
  - Calculer les fréquences et effectifs (observés) : nf_i

- Standardiser la distribution empirique (n sujets) en 9 classes (Stanines linéaires)
- Calculer les effectifs théoriques np_i imposées par la loi normale, c'est-à-dire en multipliant n par les fréquences théoriques correspondant aux 9 classes, soit : p_i = . 4/.7/.12/.17/.20/.17/.12/.7/.4.
- Calculer la quantité : $d^2 = \sum_{1}^{9} \frac{(nf_i - np_i)^2}{np_i}$
- Comparer cette quantité au seuil déterminé par un domaine de rejet de 5% dans la distribution de χ_8^2 , à savoir 15.5.

• *Remarque* : le fait de rejeter l'hypothèse nulle de normalité n'oblige pas nécessairement à normaliser la distribution, tout dépend de la caractéristique mesurée. S'il s'agit d'un trait de personnalité, rien n'indique que celui-ci doive être normal dans la population.

D.3. Test d'ajustement à une distribution théorique discrète (uniforme)

Prenons prétexte d'une pseudo-vérification empirique de l'astrologie pour introduire à la méthode utilisée par Gauquelin, le **test d'ajustement** dû à K. Pearson (technique largement utilisée en sciences humaines et naturelles depuis le début du siècle). Comme il s'agit d'un exercice académique, la maigreur de l'effectif n'a pas beaucoup d'importance, ce qui ne nous empêchera pas de discuter des caractéristiques de l'expérience, en conclusion…

1. Problématique :

On s'intéresse à vérifier la pertinence d'une typologie en matière de prédiction de l'orientation professionnelle.

2. Plan expérimental :

Vérification de la **validité concourante**, avec comme critère : la filière de formation, et comme « prédicteur » : le signe astrologique de naissance.

Plan simplifié : on ne considère qu'une filière (filière universitaire en psychologie) et on interroge un échantillon (n=135) au sujet du signe de naissance des individus qui le constituent.

On prend note que l'échantillon n'est pas représentatif des étudiants en psychologie en général, mais seulement de ceux qui suivent certaines formations spécialisées. Le tirage n'étant pas aléatoire, l'échantillon est dit « de **convenance** », il ne s'agit donc pas d'un plan expérimental au sens propre.

3. *Modèle théorique attendu :*

L'hypothèse « nulle » (appelée en général H_o) postule que les signes du zodiaque sont *distribués aléatoirement* dans l'échantillon (aucune influence du signe sur l'orientation).

On postule donc un *modèle uniforme* de la distribution des probabilités d'attribution d'un signe à un sujet, quel qu'il soit. On est donc conscient que ce modèle ne correspond peut-être pas à la distribution théorique réelle des signes de naissance dans la population globale.

Les fréquences attendues sont, dans ce modèle : $p_{1,} ..., p_{12} = 1/12$
Ce modèle est celui de l'urne remplie de boules de douze couleurs différentes, en nombre égal, ou du … « dé à douze faces ».

4. *Expérience aléatoire (statistique)*

Tirer aléatoirement n boules d'une urne dont le contenu est en principe (hypothèse nulle) conforme au modèle d'uniformité décrit ci-dessus.

Ou autrement dit : interroger les n sujet de l'échantillon à propos de leur signe de naissance.

Compter les occurrences des signes astrologiques et calculer les fréquences f_1 ... $f_i \ldots f_{12}$, réalisations des variables aléatoires : F_1 ... $F_i \ldots F_{12}$.

On note que les p_i sont des *nombres*, alors que (avant que l'expérience aie réellement lieu) les F_i constituent *des variables aléatoires*.

En effet, pour F_i, on peut imaginer autant de valeurs f_i qu'il existe de manières différentes de tirer notre échantillon de n individus dans un réservoir (population) en principe illimité.

Si l'on s'intéresse maintenant à la *distance* entre ce que l'on observe et ce que l'on attend (conformément au modèle), on s'intéressera nécessairement à la *différence* entre les quantités nf_i (effectifs observés) et np_i (effectifs attendus sous H_o).

les $nf_i - np_i$ sont donc des *variables*, puisque les f_i le sont.

(Rappelons que toutes les valeurs p_i sont égales (= 1/12)

L'idée intuitive de *distance globale entre les distributions observée et attendue* implique une *sommation* de ces différences, effectif par effectif : $\sum_1^{12}(nf_i - np_i)$.

Cependant, la possibilité de termes négatifs ne permet pas d'établir un lien nécessaire entre cette somme et la distance recherchée, c'est pourquoi on effectuera la sommation sur des *carrés de différences*.

Finalement, on obtient une *estimation* de la distance[13] entre les deux distributions (observée et théorique) en rapportant les carrés des différences à l'effectif théorique correspondant, cette estimation est appelée D^2.

Le résultat du calcul sera donc une « distance carrée » désignée par le symbole :

$$d^2 = \sum_{1}^{12} \frac{(nf_i - np_i)^2}{np_i}$$

Cette quantité d^2(réalisation de variable de décision) est l'*estimateur* de la distance réelle Δ^2 entre la distribution de probabilités d'appartenance à un signe du zodiaque dans la population de psychologues, et la distribution uniforme du « modèle ».

5. *La question décisive :*

Cette distance réelle Δ^2, estimée par D^2 peut-elle être considérée (intuitivement) comme nulle (non-rejet de l'hypothèse nulle), ou différente de zéro, auquel cas il faudrait envisager un effet des astres sur les choix en matière d'orientation professionnelle ?

Pour répondre à cette question, il faut *tester l'hypothèse nulle* en regard d'*une* expérience, à défaut de mieux.

Mais auparavant, il faut examiner le comportement de D^2, dont notre unique expérience va fournir *une* réalisation d^2 (un nombre, cette fois-ci) de la variable D^2, appelée aussi *variable de décision*.

Or, il se trouve que la quantité variable D^2, exprimant la distance entre deux distributions, l'une observée et l'autre théorique, a une distribution de probabilité connue, en particulier si l'hypothèse nulle est vraie.

Cette distribution a été calculée par les statisticiens et porte le nom de *distribution du chi-carré*. Ces distributions forment une famille, il y en a une différente pour chaque cas, selon le nombre de catégories sur lesquelles on calcule D^2.

13. On remarquera en passant que dans ce raisonnement, la notion intuitive de « distance » ne correspond pas à la notion algébrique, en effet, même si dans le cas « H_0 vraie » la distance entre le profil estimé et le profil théorique est, intuitivement parlant, nulle, il se trouve que l'espérance mathématique de l'estimateur de cette distance n'est pas nulle (en fait égale à n pour une loi chi-carré n), du simple fait que cet estimateur ne peut pas prendre des valeurs négatives, étant donné qu'il est une somme de carrés!

En particulier, la variable de décision D^2 suit une distribution de chi-carré à 11 degrés de liberté (12 – 1 cases peuvent être remplies librement, connaissant le total = n).

On peut donc consulter des tables pour examiner le comportement attendu des valeurs de D^2, on voit immédiatement que *si H_o est vraie* :

- D^2 peut varier de 0 à l'infini, mais la valeur théorique la plus probable (l'espérance mathématique) de χ^2_{11} est **11**, et non pas zéro, contrairement à l'attente intuitive (*cf.* note 10).
- On voit aussi que D^2 ne dépassera la valeur de 17.27 que dans 10% des cas, et la valeur 19.67 dans 5% des cas.

Le cadre théorique étant maintenant parfaitement décrit, on peut maintenant effectuer un test d'ajustement, à l'aide d'*une* expérimentation pratique.

6. *Test du « goodness of fit » (dû à K. Pearson, env. 1900)*

Connaissant la distribution des valeurs attendues de D^2 dans le cas de H_o vraie, adoptons l'attitude suivante : « tirons » (ou contentons-nous des sujets à disposition) un échantillon de 135 individus et calculons pour cette expérience particulière la valeur de l'estimateur D^2, et comparons ensuite cette valeur à celles attendues.

Expérience faite, nous trouvons d^2 (*réalisation* de D^2) = 16.02.

Que penser alors de notre hypothèse nulle ?

- Si le modèle est vrai, on attend une valeur proche de 11, ou du moins pas trop éloignée...
- On imagine bien que si on avait trouvé 56 par exemple, on ne pourrait plus croire que le modèle est acceptable, et on serait forcé d'admettre que *cette expérience* pourrait attester de l'effet des astres sur la profession envisagée.
- Notre d^2 est le centile 87 de la distribution attendue, il est donc « assez rare » de tomber sur un tel échantillon (13% des cas), en supposant que le modèle soit « vrai ».

Nous pouvons en conclure que l'effet « astral » observé n'est *peut-être* pas seulement dû au hasard, notre confiance dans le modèle d'uniformité est un peu diminuée, mais seule une nouvelle expérience (au moins) nous fixera plus précisément à ce sujet.

Ce type de raisonnement aurait sans doute été celui de *Fisher* (*cf.* article actualités psychologiques).

Mais supposons maintenant que cette expérience serve de base à une *décision*, par exemple de préconiser aux conseillers d'orientation de diriger vers la psychologie tous les sujets nés sous le signe du taureau et intéressés par les sciences de l'esprit, mais indécis (…)

Il s'agit alors de définir une *règle de décision*. Formellement, il s'agit donc de choisir entre deux hypothèses alternatives :

- H_o : Les signes de naissance sont distribués aléatoirement dans la population des étudiants en psychologie.
- H_1 : Les signes de naissance ne sont pas distribués aléatoirement dans la population, auquel cas la connaissance du signe de naissance serait *prédicteur* d'une certaine forme d'intérêt pour les sciences humaines.

Précisons, en principe avant l'expérience, la valeur seuil (ou critique) que D^2 ne devrait pas dépasser, auquel cas on décidera que l'hypothèse nulle doit être rejetée au profit d'une hypothèse alternative H_1 : les signes de naissance des sujets de la population de psychologues ne sont pas distribués aléatoirement.

Ce seuil sera déterminé par la probabilité pour D^2 de « tomber » dans des valeurs extrêmes, excluons donc le 5% des valeurs de ce type et attribuons-les à une *zone de rejet* de H_o, donc à celle de l'adoption de H_1. Dans notre cas, la valeur de 19.67 constitue la valeur seuil recherchée (*cf.* table).

La règle de décision prend donc la forme suivante :

Si pour une expérience donnée, le D^2 calculé est inférieur à 19.67, rien ne nous « signifie » qu'il faut rejeter l'hypothèse nulle, alors que si cette valeur dépassait le seuil fixé, on interprétera ce « signe » comme un déni expérimental de H_o qui devrait alors être rejetée au profit d'une hypothèse alternative, avec les conséquences pratiques qui s'en suivent.

En ce qui concerne notre expérience, l'hypothèse nulle n'est donc pas rejetée, la valeur légèrement excessive de la variable D^2 peut être attribuée au seul aléa d'échantillonnage. En d'autre termes, cette expérience ne nous permet pas de rejeter le modèle de répartition uniforme des signes dans la population parente (notre d^2 – réalisation de la variable de décision D^2 – n'est pas « *significatif* »).

Ce type de raisonnement, très pragmatique, date des alentours de 1933 et est dû à *J. Neyman* et *Egon Pearson*, fils de Karl.

Rappelons les implications pratiques des deux attitudes : l'attitude fishérienne renvoie à une conception épistémique de l'induction : l'expérience permet d'accroître notre

connaissance de la réalité en précisant la fiabilité de la vérité de certaines hypothèses. Le problème du *risque d'erreur* n'existe donc pas chez *Fisher*.

Il n'en va pas de même si l'on applique une règle de décision, si une hypothèse est préférée à une autre, et que ce choix a des conséquences pratiques, alors il est nécessaire de *quantifier le risque* que l'on court lorsqu'on agit de la sorte. Dans notre cas, si l'on déclare qu'une valeur observée D^2 comprise dans le 5% des valeurs extrêmes de la distribution de chi-carré « signe » le rejet de H_o, alors il faut s'attendre à se tromper 5 fois sur 100 expériences, puisqu'il est clair, d'après la distribution théorique, que 5 expériences sur 100 fournissent de telles valeurs, *même si l'hypothèse nulle est vraie*!

Dans cet exemple, on supposait qu'il n'est pas trop grave d'orienter des gens déjà intéressés par les sciences psychologiques vers ce type d'orientation, sur la seule base du critère « signe astrologique », d'où le choix d'un seuil peu exigeant.

Dans la vision fréquentiste des probabilités, le risque de première espèce (rejeter H_o alors qu'elle est en fait vraie) est donc égal à la probabilité de trouver une valeur supérieure au seuil (valeur de chi-carré) fixé. Dans le langage méthodologique actuel, on appelle seuil aussi bien la valeur donnée par la table (19.67) que la probabilité cumulée que chi-carré dépasse cette valeur (5%). Dans cette conception, la probabilité d'un risque est simplement associée à une fréquence.

Quand au *risque de seconde espèce* (ne pas rejeter H_o, alors que H_1 est vraie, ce qui pourrait être le cas ici…) il ne peut être quantifié lorsque H_1 est simplement complémentaire à H_o. Il n'existe en effet pas de distribution théorique pour une hypothèse composite de ce genre.

- *Conclusions*

On peut s'interroger sur les enseignements de cette expérience :

- Du point de vue du progrès de la connaissance, on reste sur sa faim, une seule expérience ne permet pas de mettre une hypothèse en doute, surtout si elle est bien ancrée dans la rationalité, comme le modèle d'uniformité des naissances.
- On sait d'autre part que même en l'absence de toute influence astrale, il est probable que cette hypothèse théorique ne corresponde pas à la réalité. Une plus ample documentation est nécessaire pour juger du phénomène de l'irrégularité saisonnière des naissances. Ce fait n'aide pas à clarifier la situation…
- Et finalement, comment distinguer une éventuelle influence des astres de celle d'autres facteurs saisonniers ? Si H_o avait été rejetée, aurions nous réellement tenu un argument pour l'astrologie ? Certainement pas, tout au plus une piste d'investigation à explorer de manière plus sérieuse (plus grands échantillons, meilleure représentativité, etc…).

On a donc appris bien peu de choses, et c'est ce qui peut expliquer la relative indifférence du public aux arguments de Gauquelin. Même s'il peut exhiber quelques résultats « significatifs », ceux-ci ne peuvent nous convaincre de la réalité de l'influence des astres sur l'orientation professionnelle.

En résumé : test d'ajustement à une distribution théorique discrète - mode d'emploi

- *Conditions d'utilisation* : pas plus d'un quart des effectifs théoriques ne doivent être inférieurs à 5. Les individus doivent être tirés au hasard.
- *Procédure à suivre* :
  - Calculer les c fréquences et effectifs (observés) : nf_i
  - Calculer les c effectifs théoriques np_i imposés par le modèle.
  - Calculer la quantité : $d^2 = \sum_1^c \frac{(nf_i - np_i)^2}{np_i}$
  - Comparer cette quantité au seuil déterminé par un domaine de rejet de 5% (ou 1%) dans la distribution de χ^2_{c-1}

D.4. Test d'ajustement à une proportion théorique

1. Problématique :

On se souvient des efforts de Kretschmer qui cherchait un lien entre la constitution physiques des êtres humains et leurs caractéristiques tempéramentales (*cf.* cours évaluation psychologique). Cherchons plus précisément à savoir si, pour une femme, le fait d'être cataloguée « schizothyme » par la théorie de *Kretschmer* implique qu'elle appartienne plus souvent au type physique « leptosome », plutôt qu'à tout autre.)

TABLEAU 4. Kretschmer : table de contingences entre morpho-types et psycho-types « normaux » (femmes)

	Schizothyme	Cyclothyme	Total
Pycnique	29	202	231
Leptosome	432	86	518
Athlétique	101	14 (O_{32})	115
Total	562	302	864

2. Plan expérimental :

Imaginons l'expérience aléatoire : tirer un échantillon de n individus (femmes, par exemple) déclarés « schizothymes » et demander à un expert de les ranger selon leur type physique d'après la méthode de *Kretschmer*.

3. Modèle théorique attendu :

Décrire un modèle d'indépendance (*i.e.* postuler que *Kretschmer* a tort) suppose un peu de réflexion. Comme dans le cas de la répartition des signes astrologiques, on pourrait postuler un *modèle uniforme* de répartition des probabilités, si bien que si le classement des sujets se ferait en trois types, selon des probabilité égales, à savoir 1/3 = .33. Autrement dit, dans un tel modèle, la probabilité d'être classée « leptosome » pour une femme schizothyme est de .33 ; contre .66 d'être classée autrement.

Cependant, les chiffres fournis par *Kretschmer* ne nous permettent pas de postuler un modèle équiprobable. Si l'on admet que l'échantillon étudié constitue un échantillon plus ou moins aléatoire, il est clair que les « leptosomes » sont plus nombreux dans la population courante que les individus des deux autres types. Plus précisément, en ce qui concerne les femmes, *Kretschmer* a observé 518 types leptosomes sur 864 femmes, ce qui donne une probabilité théorique (à défaut de mieux, car on est obligé de croire les chiffres de l'auteur) de .6. Finalement, pour une femme classée « schizothyme », la probabilité de ne pas être « leptosome » est donc de .4. Résumons donc H_o :

- p = .60 est, pour une femme, la *probabilité théorique* d'être classée « leptosome » *si* le modèle d'indépendance est vrai – et si Kretschmer classe correctement les types physiques !

- et : (1 – p) = .40 , celle d'être classée autrement.

4. Expérience aléatoire :

Soit l'expérience aléatoire abstraite : *tirer 562 sujets*[14] *au hasard dans une population de femmes déclarées appartenir au type psychique « schizothyme », et noter leurs types physiques*. Les sujets sont classés en deux catégories : les « leptosome » et les « autre ».

Nous désignerons par F la proportion de « leptosomes », or cette quantité F est une variable (tant que l'expérience n'a pas réellement eu lieu) dont on aimerait bien connaître le comportement, si le modèle est effectivement valable dans la population dont l'échantillon a été tiré. On sait maintenant que si le modèle est vrai dans la population d'où l'échantillon a été tiré, et pour n assez grand (minimum 30), la distribution échantillonnale des valeurs de F suit une loi de probabilité gaussienne, d'espérance p, et d'écart-type : $\sqrt{\frac{(1-p)\cdot p}{n}}$.

14. *Cf.* le total marginal « colonne » correspondant au type considéré, dans le tableau des effectifs observés.

Décidons que si le modèle est valable, alors la valeur numérique f observée dans une expérience particulière devrait se trouver contenue dans un intervalle de confiance aux limites bien définies, construit autour de la valeur théorique p (p étant l'espérance mathématique de F).

Définissons les bornes de cet intervalle. Si F est distribuée de manière gaussienne, alors elle se comporte comme les valeurs standardisées **u**, telles que décrites dans une table de la loi normale.

On y voit que, par exemple, la valeur de ±1.96 (on peut arrondir à 2 pour les calculs rapides) marque la limite des 2.5% inférieurs et des 2.5% supérieurs. On calcule ainsi un intervalle ayant 95% de chances d'« accueillir » notre valeur expérimentale, si H_o est vraie.

Comme notre variable « F » a une moyenne (*espérance*) p = 0,60 et un sigma (écart-type) de $\sqrt{\frac{(1-0.60)\cdot 0.60}{562}} = 0.02$, en procédant à l'opération inverse de la standardisation, on peut facilement calculer les bornes d'un intervalle à 95% construit autour de 0,60 :

- Borne supérieure : $(0.60) + (0,02 \cdot 1.96) = 0.60 + 0.04 = 0.64$
- Borne inférieure : $(0.60) - (0,02 \cdot 1.96) = 0.60 - 0.04 = 0.56$

5. La question décisive :

Notre intervalle contiendra-t-il la valeur observée lors d'une seule expérience ? *Expérience faite,* nous trouvons f = 432/562 = .768 , valeur qui ne tombe pas dans l'intervalle de confiance défini ci-dessus.

6. Que penser après cette expérience?

Le test aboutit donc à un relatif discrédit de l'hypothèse nulle (attitude de *Fisher*), ou à son rejet pur et simple au seuil 5%, si l'on suit une règle de décision à la Neyman-Pearson (en excluant les 2.5% extrêmes de notre intervalle de confiance, on fixait un seuil implicite à 5%). À la lumière de cette nouvelle expérience, nous sommes forcés de croire que Kretschmer a peut-être raison : les individus schizothymes sont plutôt du type longiligne…

Mais nous savons aujourd'hui comment ces données ont été biaisées par les *a priori* du chercheur, si bien que notre conclusion est erronée : la théorie de Kretschmer n'a plus guère d'adeptes de nos jours.

En résumé : test d'ajustement à une proportion théorique - mode d'emploi

- *Conditions d'utilisation* : les individus doivent être tirés au hasard.

- *Procédure à suivre* : soit p la proportion théorique et f la proportion observée :

  - Calculer la quantité : $z = \dfrac{f - p}{\sqrt{\dfrac{(1-p)\cdot p}{n}}}$

  - Si H_0 est vraie, z suit une loi normale centrée réduite, il suffit donc de comparer cette quantité au seuil déterminé par un domaine (bilatéral) de rejet de 5% (ou 1%) dans la distribution normale, soit ± 1.96 (ou ± 2.54).

  - La quantité (p - f) peut directement être interprétée comme une taille d'effet à laquelle on peut appliquer les critères de Cohen.

- *Remarque* : les quantités np ou n(1-p) (effectifs attendus) doivent être toutes deux supérieures à 5.

D. 5. Test d'ajustement à une moyenne théorique

Ce problème a été celui de W. Gosset (Student) qui travaillait dans une brasserie de bière et était chargé de surveiller les taux de diverses substances présentes dans les tonneaux en fermentation. Dérangé par l'odeur des tonneaux, Gosset chercha à vérifier – sans avoir à ouvrir tous les tonneaux – si les taux X d'une certaine substance avaient bien pour moyenne une valeur théorique μ.

Connaissant le théorème central limite, Gosset savait que le taux moyen M mesuré sur un échantillon de taille n suivait une loi normale d'espérance μ et d'écart-type $\dfrac{\sigma}{\sqrt{n}}$, si H_0 : l'échantillon est tiré d'une population dans laquelle m=μ. Il lui aurait alors été facile de tester la quantité standardisée : $Z = \dfrac{M - \mu}{\dfrac{\sigma}{\sqrt{n}}}$ qui suit une loi normale centrée réduite si H_0 (<=> l'échantillon est tiré d'une population dans laquelle la moyenne du caractère mesuré est μ) est vraie.

Pour Gosset, l'ennui résidait dans le fait que s'il connaissait la moyenne μ des taux X dans la population de tonneaux, il en ignorait par contre la variance σ^2 ! Il lui a donc fallu l'*estimer*.

C'est naturellement la variance S^2 de l'échantillon qui lui servit d'estimation de σ^2. Mais du coup, le théorème central limite ne s'appliquait plus et Gosset dut s'adresser à des mathématiciens pour connaître la distribution théorique de M, si H_0 est vraie.

Il apparut que M suivait une distribution proche de la normale, mais néanmoins différente, surtout dans les cas où l'échantillon était petit. Cette nouvelle distribution fut appelée par Gosset « distribution du t de Student ». Ainsi, si σ est estimée par S, écart-type de l'échantillon, la distribution échantillonnale de M suit une loi **t**, d'espérance μ et d'écart-type $\frac{S}{\sqrt{n}}$. Il était dès lors possible de tabuler la distribution de la variable t standardisée, pour divers degrés de liberté.

En résumé : test d'ajustement à une moyenne théorique - mode d'emploi

- *Conditions d'utilisation* : les individus doivent être tirés au hasard et le caractère X de moyenne μ doit être normalement distribué.
- *Procédure à suivre* : soit μ la moyenne théorique, m la moyenne l'échantillon de taille n et s son écart-type.

  - Calculer la quantité : $t = \dfrac{m - \mu}{\frac{s}{\sqrt{n}}}$
  - Si H_o est vraie, t suit une loi de t de Student à n-1 degrés de liberté, il suffit donc de comparer cette quantité au seuil déterminé par un domaine de rejet de 5% (ou 1%) dans la distribution de t de la table.
  - La quantité : $d = \dfrac{m - \mu}{s}$ est la taille de l'effet dû à l'appartenance au groupe expérimental *vs* théorique.
- *Remarque* : la statistique t est robuste et supporte bien la violation de la règle de normalité du caractère X dans la population. Le problème est plus délicat pour les tests unilatéraux.

Exemple (sans analyse de puissance *a priori*) **:**

On mesure sur un échantillon de 25 enfants une moyenne de 113.64 à une épreuve de Q. I. L'écart-type est de 12.4. On désire savoir si cet échantillon peut être considéré comme tiré d'une population générale dans laquelle la moyenne est de 100.

H_o : l'échantillon est tiré d'une population où $\mu = 100$

on calcule : $t = \dfrac{m - \mu}{\dfrac{s}{\sqrt{n}}} = \dfrac{113.64 - 100}{\dfrac{12.4}{\sqrt{25}}} = 5.5$

la taille de l'effet est égale à : $\dfrac{113.64 - 100}{12.4}$ = environ 1, la différence observée est donc très importante.

Cette quantité peut être comparée à la valeur de $t_{[24]}$ au seuil 5% qui est 2.064, si H_o est vraie. On constate que notre valeur dépasse largement ce seuil, ce qui nous incite à rejeter l'H_o et à conclure provisoirement (avec un risque de première espèce de 5%) que notre échantillon provient d'une population particulière, dans laquelle le Q.I. moyen semble supérieur à celui de la population générale. Comme nous n'avons pas émis d'hypothèse alternative, nous n'avons pas d'autre estimation de ce niveau supérieur que celle fournie par notre échantillon, à savoir 113.64. Toute la description de cette population particulière reste donc à faire.

E. Tests d'indépendance

Précisons tout d'abord qu'il ne faudrait pas considérer les tests d'indépendance comme une classe de tests complètement distincts des tests d'ajustement ; en fait, ils n'en constituent qu'une sous-catégorie, à savoir celle des *test d'ajustement à un modèle particulier : celui de l'indépendance*. Par exemple, nous avons déjà vu qu'un test d'ajustement à une distribution *uniforme* revenait en fait à établir l'indépendance entre une variable catégorielle (signe astrologique) et une autre variable catégorielle (appartenance à une filière de formation *vs* non appartenance). Le cas est particulièrement évident avec le coefficient de corrélation qui permet de tester l'ajustement à une valeur déterminée, pratiquement toujours zéro, qui représente justement l'indépendance entre deux variables.

Les tests d'indépendance constituent donc une sous-classe restreinte comprise dans l'ensemble des tests d'ajustement. Si les tests d'indépendance sont si universellement connus et utilisés, c'est que le modèle de l'indépendance, généralement formalisé sous l'appellation « H zéro » est de loin le plus facile à tester à l'aide des techniques inventées par Pearson ou Fisher, que l'on soit en présence de variables numériques, ordinales ou catégorielles.

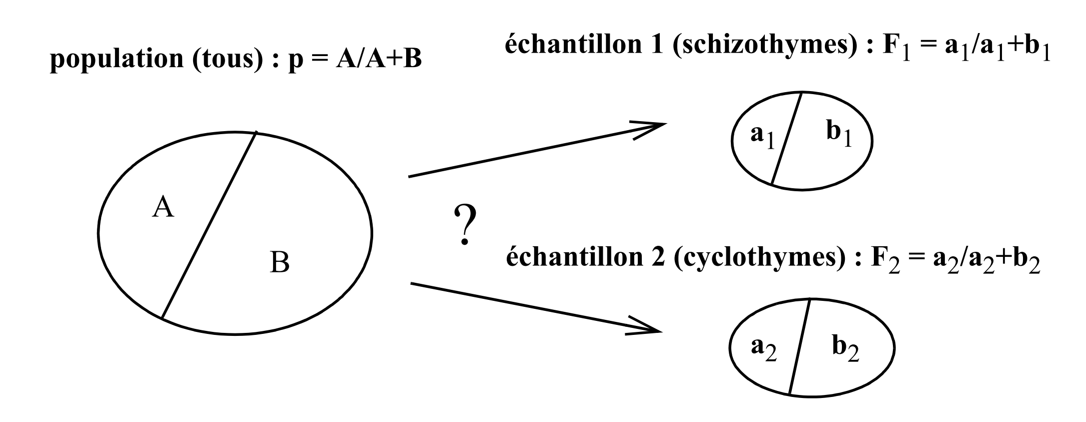
E.1. Indépendance entre deux variables catégorielles

E.1.1. Cas 1 : test d'indépendance entre deux variables catégorielles dichotomiques - *Comparaison de deux proportions observées*

Nous avons déjà abordé plus haut le problème de la comparaison d'une proportion observée avec une proportion théorique ; le problème revenait à se donner les moyens de décider si l'échantillon dans lequel on avait mesuré la proportion observée f pouvait être considéré comme tiré d'une population dans laquelle la proportion théorique était égale à une valeur p donnée.

On rencontre souvent une situation différente : la proportion p de la population est inconnue, mais on dispose de deux échantillons dont on se demande s'ils sont tirés de la même population. On peut par exemple mesurer la proportion de « leptosomes » dans un premier échantillon d'hommes schizothymes (932 sur 1258 selon Kretschmer), puis mesurer de la même manière la proportion d'hommes leptosomes et cyclothymes (183 sur 756). Si ces deux proportions diffèrent de manière « significative », on en déduira qu'il existe un lien entre le type physique et le type psychique d'un individu, c'est pourquoi les techniques de comparaison de proportions peuvent être considérées comme des tests d'indépendance entre variables catégorielles. Dans notre exemple les deux variables catégorielles sont le *type physique* et le *type psychique*. On traite ce genre de problème en testant une hypothèse nulle d'indépendance : les deux groupes proviennent d'une même population, à savoir celle des hommes en général (en supposant que la distinction entre schizothyme et cyclothyme couvre l'entier de la population).

Cette H_o peut se représenter de la manière suivante :



- et si elle est vraie, F_1 et F_2 sont des variables aléatoires d'échantillon, d'espérance p et d'écart-type : $\sqrt{\frac{(1-p)\cdot p}{n_1}}$ pour l'échantillon 1 (de taille n_1) ; et d'écart-type : $\sqrt{\frac{(1-p)\cdot p}{n_2}}$ pour l'échantillon 2 (taille n_2).

- Le problème est que cette fois-ci nous ne connaissons pas p ! Il s'agit donc de l'estimer à partir des proportions observées, en les pondérant en fonction de la taille de l'échantillon :

$$\hat{p} = \frac{n_1 \cdot F_1 + n_2 \cdot F_2}{n_1 + n_2}$$

F_1 suit donc une loi normale d'espérance $\hat{p}$ et d'écart-type $\sqrt{\frac{(1-\hat{p}) \cdot \hat{p}}{n_1}}$;

et F_2 suit une loi normale d'espérance $\hat{p}$ et d'écart-type $\sqrt{\frac{(1-\hat{p}) \cdot \hat{p}}{n_2}}$.

- On s'intéresse maintenant à la différence de ces deux quantités, car il est clair que si H_o est vraie, l'espérance de cette différence sera zéro, alors que son écart-type sera égal à la racine carrée de la somme des variances[15]. Or, comme :

$$\sqrt{\frac{(1-\hat{p}) \cdot \hat{p}}{n_1} + \frac{(1-\hat{p}) \cdot \hat{p}}{n_2}} = \sqrt{\langle [1-\hat{p}] \cdot \hat{p} \rangle \cdot \langle \frac{1}{n_1} + \frac{1}{n_2} \rangle} = \sqrt{(1-\hat{p}) \cdot \hat{p}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

il s'en suit que la variable $F_1 - F_2$ suit une loi normale d'espérance 0 et d'écart-type : $\sqrt{(1-\hat{p}) \cdot \hat{p}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

- Finalement, si H_o est vraie, la variable standard :

$$z = \frac{F_1 - F_2}{\sqrt{(1-\hat{p}) \cdot \hat{p}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

suit une loi normale de moyenne 0 et d'écart-type 1, ce qui permet de consulter la table de répartition de u pour trouver les seuils correspondants aux domaines critiques α = 5% ou 1 %.

- Par exemple, au seuil 5%, z ne doit pas dépasser ± 1.96 (test bilatéral), sinon H_o devra être rejetée.

- Si l'on reprend l'exemple du début, la proportion de « leptosome et schizothyme » est de 932/1258 = .74 ; et la proportion de « leptosome et cyclothyme » est de 183/756 = .24.

15. La variance d'une somme ou d'une différence de deux variables *indépendantes* est égale à la somme des variances : $\sigma^2_{(X \pm Y)} = \sigma^2_{(X)} + \sigma^2_{(Y)}$.
C'est pourquoi : $\sigma_{\langle X \pm Y \rangle} = \sqrt{\sigma_X^2 + \sigma_Y^2}$

On calcule donc tout d'abord la quantité :

$$\hat{p} = \frac{1258 \cdot 0,74 + 756 \cdot 0,24}{1258 + 756} = 0,55$$

puis :

$$z = \frac{0,74 - 0,24}{\sqrt{(1 - 0,55) \cdot 0,55} \cdot \sqrt{\frac{1}{1258} + \frac{1}{756}}} = 21,8$$

valeur qui excède largement le seuil de 1,96. L'hypothèse nulle ne peut qu'être rejetée : si on en croit les chiffres de Kretschmer, le type physique et fortement lié au type psychique.

Exemple : on se demande si le comportement tabagique des adolescents dépend du sexe.

- On se propose donc de tester le lien entre deux variables qualitatives dichotomiques : *sexe* (modalités : F ou M), et *comportement tabagique* (modalités : fume/ne fume pas).
- On interroge 100 garçons, dont 32 disent qu'ils fument, et 80 filles dont 28 disent aussi fumer.
- L'H_o est définie comme suit : les deux échantillons sont tirés de la population générale des adolescents dans laquelle les filles et les garçons fument dans la même proportion *p* (inconnue).
- On peut disposer les données de la manière suivante :

TABLEAU 5. Table de contingences « sexe » et « comportement tabagique »

	Fume	Ne fume pas	total
Filles	28	52	80
Garçons	32	68	100
total	60	120	**180**

Un tel tableau se prête très bien à un test du chi carré (*cf.* § suivant), mais on peut aussi l'utiliser pour illustrer le point traité ici, à savoir le *test sur les proportions*.

- On calcule $f_1 = 28/80 = 7/20$ pour les filles, et $f_2 = 32/100 = 8/25$ pour les garçons, donc $f_1 - f_2 = 0.03$, et $\hat{p} = 60/180 = 0,33$.

- Et ensuite : $z = -\frac{0,03}{0,07} = -0,425$, valeur qui ne dépasse pas le seuil fixé. Cette expérience ne permet donc pas de conclure que le comportement tabagique dépend du sexe des adolescents.

E.1.2. Cas 2 : test d'indépendance entre deux variables catégorielles quelconques - *le test du « chi carré »*

Le test dit du « chi carré » est une technique permettant de juger du degré de dépendance entre deux variables catégorielles quelconques (du point de vue du nombre de modalités).

L'hypothèse nulle est toujours une hypothèse d'indépendance.

Exemple : *tentative de validation de la typologie de Kretschmer : test d'indépendance de deux variables catégorielles*.

On peut tout d'abord se contenter de remarquer que pour les groupes étudié par Kretschmer (hommes et femmes), les proportions observées sont suffisamment explicites pour corroborer sa théorie (*cf.* Tableau 1 *supra* et 2 suiv.).

Cependant, cette analyse descriptive est d'une portée limitée, sans grand intérêt pour le progrès de la connaissance. Ce qui nous intéresse en réalité, et c'est sans aucun doute également l'intention de *Kretschmer*, c'est de démontrer la validité de sa théorie pour tous les hommes ou toutes les femmes. Il faut donc adopter un point de vue inférentiel : que nous apprend cette observation particulière sur la population générale (appelée « parente » par les statisticiens) ?

TABLEAU 6. Kretschmer : table de contingences entre morpho-types et psycho-types « normaux » (hommes)

	Schizothyme	Cyclothyme	Total
Pycnique	21	547	568
Leptosome	932	183	1115
Athlétique	305	26	331
Total	1258	756	**2014**

Un tel tableau croisé (*cf.* Tableau 6), ou table de contingence (montrant des « liens »), peut être considéré de plusieurs manières, selon la portée de l'inférence envisagée :

- Point de vue global : y a-t-il, dans la population des hommes, un lien entre le type physique et le tempérament, comme le prétend l'auteur sur la base de son expérience ?

- et un autre point de vue, plus ponctuel, par exemple : Les hommes « schizothymes » (en général) appartiennent-ils bien au type « leptosome » plutôt qu'à tout autre ? Ce problème a déjà été traité (dans le cas des femmes) dans le cadre du test d'ajustement à une proportion théorique.

1. Problématique :

On s'intéresse à vérifier la pertinence d'une théorie du déterminisme corps – esprit, à propos d'une population d'hommes.

2. Plan expérimental :

Vérification de la validité concourante par « corrélation » entre type physique et psychologique. La notion de corrélation étant liée à l'approche statistique numérique, on préférera parler dans ce cas de *test d'indépendance* entre deux variables catégorielles.

Les données à disposition sont celles de l'auteur de la théorie : on doit faire confiance… Ce qui signifie qu'on ne sait pas précisément comment *Kretschmer* a choisi ses sujets d'expérience.

Les deux variables catégorielles (ou nominales) en question sont : *Type physique* (3 catégories) et *Type psychique* (2 catégories). Le tableau de contingence a donc 2 x 3 = 6 cases.

On appelle *totaux marginaux* les sommes par ligne et par colonne.

On appelle *total général* ou *effectif* de l'échantillon la somme des totaux marginaux-lignes (ou colonnes).

3 Modèle théorique attendu :

Il faut admettre comme un principe méthodologique que les modèles théoriques postulent souvent l'*indépendance* des variables en question pour des raisons de simplicité des calculs. Ce postulat revient à dire que le hasard seul explique les différences de répartition dans le tableau (en fonction des totaux marginaux, bien entendu).

Mais on peut évidemment aussi tester des modèles de dépendance plus complexes, il s'agit alors plutôt de mettre en place une règle de décision permettant de choisir entre deux ou plusieurs modèles.

Dans notre cas, et en fonction de la problématique définie ci-dessus, on postulera un modèle d'indépendance entre les types physiques et psychologiques (dans la population considérée).

Une telle hypothèse « nulle » permet de calculer un tableau d'*effectifs attendus* (*i.e.* ce que l'on devrait observer le plus probablement si le modèle d'indépendance est vrai...) faciles à calculer :

TABLEAU 7. Effectifs attendus (e_{ij}) en cas d'indépendance entre morpho - et psychotypes (hommes)

	Schizothyme	Cyclothyme	Total
Pycnique	*354.8*	*213.2*	568
Leptosome	*696.5*	*418.5*	1115
Athlétique	*206.7*	*124.3*	331
Total	1258	756	**2014**

Pour trouver, par exemple, l'effectif attendu de la case « Pycnique et Schizothyme », on multiplie la probabilité d'être de type pycnique (568/2014) par celle d'être Schizothyme (1258/2014). Ces deux « événements » étant supposés (par H_o) être indépendants, le produit obtenu représente bien la probabilité de figurer dans la case « Pycnique *et* Schizothyme », à savoir $\frac{568 \cdot 1258}{2014 \cdot 2014}$. Sachant que l'effectif total est de 2014, l'*effectif attendu* dans la case en question sera donc : $\frac{568 \cdot 1258 \cdot 2014}{2014 \cdot 2014} = \frac{568 \cdot 1258}{2014} = 354.8$

4. Expérience aléatoire :

Comme dans un test d'ajustement à une distribution théorique discrète, on va s'intéresser à la « distance » entre des effectifs *observé*s lors d'une expérience, par exemple celle de *Kretschmer*, et les effectifs *attendus* si le modèle d'indépendance est vrai.

On appellera « O_{ij} » (i indique les lignes et j les colonnes), les effectifs *observés* lors d'une expérience aléatoire du type « tirer 2014 hommes au hasard et noter leurs types physique et psychologique » ; et « e_{ij} » les effectifs *attendus* (*expected*) si le modèle est vrai.

- Les e_{ij} sont des nombres (*cf.* Tableau 7)
- Les O_{ij} sont des variables, puisqu'on peut imaginer autant d'expériences aléatoires (du type décrit ci-dessus) que l'on veut, en respectant toutefois l'effectif de 2014.

Les quantités $O_{ij} - e_{ij}$ (différences case par case) sont donc aussi des variables, de même les carrés de ces quantités, et de même encore les quantités $(O_{ij} - e_{ij})^2/e_{ij}$.

Finalement, la quantité :

$$D^2 = \sum_1^3 \sum_1^2 \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

… est aussi une variable (dite « de décision ») dont on connaît la distribution théorique dans le cas où H_0 est vraie. Dans notre cas, la quantité variable D^2 suit une loi de chi carré à 2 degrés de liberté[16] [2 = (3-1)(2-1)].

Cette quantité (variable) D^2 est l'*estimateur* de la distance réelle Δ^2 entre les distributions d'effectifs observés et théoriques. Intuitivement, on s'attend à observer une distance proche de zéro, si notre échantillon de 2014 hommes est bien tiré d'une population dans laquelle le modèle d'indépendance est vrai (c'est à dire une population dans laquelle la théorie de *Kretschmer* ne classerait pas mieux les individus que ne le ferait le seul hasard).

5. *La question décisive :*

La distance D^2 observée lors d'une expérience n'est-elle pas trop éloignée de ce que l'on attend, si le modèle est vrai ? Autrement dit, la non-coïncidence des tableaux « observé » et « attendu » peut-elle être attribuée au seul aléa de l'échantillonnage, ou doit-on admettre qu'il existe bien, dans la population parente, un lien entre les variables étudiées ?

Pour répondre à cette question, il faut tester l'hypothèse nulle d'indépendance en regard d'une expérience.

Mais auparavant, il faut examiner le comportement de D^2, dont notre unique expérience va fournir *une* réalisation $\mathbf{d}^2$ (un nombre, cette fois-ci).

Nous avons vu plus haut que D^2 suit une loi de chi carré à 2 degrés de liberté ($\chi^2_{[2]}$). On peut donc se fixer quelques repères, grâce à la table du même nom. On y découvre que, si H_0 est vraie…

- D^2 peut en principe varier de 0 à l'infini, mais la valeur la plus probable est 2 (espérance mathématique[17] de $\chi^2_{[2]}$).
- On voit aussi que D^2 ne dépassera la valeur de 4.6 que dans 10% des cas, et la valeur 6 dans 5% des cas.

Cela étant connu, on peut maintenant effectuer un test à l'aide d'*une* expérimentation (en l'occurrence celle de *Kretschmer*).

6. *Test d'indépendance ou « test du chi carré »*

Expérience faite, nous trouvons d^2 = 1173, qui dépasse de loin toutes les valeurs « critiques » usuelles, qu'elles soient définies par des seuils de 5%, de 1% ou moins.

16. Dans un tableau 3 x 2 dont les effectifs marginaux sont fixés, il n'y a que *deux* cases sur six dont on puisse décider librement de l'effectif.

17. On veillera à ne pas confondre l'*espérance* mathématique d'une variable, c'est à dire sa moyenne, et la valeur *attendue* (= théorique) d'un *paramètre*.

On dit que cette valeur est « significative » car elle peut être interprétée comme un « signe » de l'éventuelle non-validité du modèle[18].

(Note : les logiciels anglo-saxons appellent cette valeur directement « chi square », convention que certains statisticiens déplorent, comme d'ailleurs la dénomination de « test du chi carré »…)

Que penser après cette expérience ?

On pouvait espérer une valeur pas trop éloignée de 2, en supposant le modèle d'indépendance vrai. La valeur trouvée (1173) entame donc très sérieusement notre conviction en sa validité (attitude « épistémique » de *Fisher*).

Si l'on avait appliqué une règle de décision au seuil 5% on aurait rejeté H_o dès que d^2 dépasse 6. Le rejet de H_o ne semble donc pas poser trop de problèmes… mais la théorie fréquentiste de Neyman & Pearson nous rappelle que cette assurance est trompeuse, car le risque d'erreur reste par définition égal au seuil fixé : 5%!

Rappelons enfin que l'histoire a montré que malgré l'évidence que l'on pourrait tirer de ces chiffres, la théorie de Kretschmer n'est pratiquement plus admise aujourd'hui : si le test est correct, les données sont, quant à elles, extrêmement biaisées et sans valeur scientifique. Dans un tel cas, les traitements les plus complexes n'aboutissent qu'à des résultats non pertinents.

En résumé : test d'indépendance entre deux variables catégorielles ; « *test du chi carré* » - mode d'emploi

- *Conditions d'utilisation* : pas plus d'un quart des effectifs théoriques ne doivent être inférieurs à 5. Les individus doivent être suffisamment nombreux et tirés au hasard.
- *Procédure à suivre* :
  - H_o : les variables sont indépendantes
  - Dresser le tableau des effectifs observés : o_{ij} ; ce tableau comporte r lignes et c colonnes.
  - Calculer les effectifs théoriques e_{ij} (attendus) découlant du modèle d'indépendance.
  - Calculer la quantité : $d^2 = \sum_{i}^{r}\sum_{j}^{c}\frac{(o_{ij} - e_{ij})^2}{e_{ij}}$

18. Les expressions « très significatif », « extrêmement significatif », et toutes autres fantaisies sémantiques du même genre, trop souvent rencontrées dans nombre d'articles « scientifiques », n'ont pas le sens propre défini ici, et devraient de ce fait être proscrites.

- Comparer cette quantité au seuil déterminé par un domaine de rejet de 5% (ou 1%) dans la distribution de : $\chi^2_{[(r-1)\cdot(c-1)]}$

• *Commentaires* :

- Dans les tableaux 2x2, la fréquence attendue la plus petite doit être supérieure à 10 ;
- Dans les tableaux 2x2, on peut transformer d^2 en un équivalent (phi) d'indice de corrélation ; $\varphi = \sqrt{\frac{d^2}{n}}$ Cet indice varie entre 0 et 1 *et peut être considéré comme un indicateur de taille d'effet*.
- Pour les tableaux plus grands, Cramér a proposé un indice phi (ou V) qui s'écrit : $\varphi_c = \sqrt{\frac{d^2}{n\cdot(k-1)}}$ où k est le plus petit des nombres de lignes ou de colonnes. Cet indice varie entre 0 et 1 (contrairement au coefficient de contingences) et peut être interprété comme la valeur absolue d'un coefficient de corrélation et *il peut donc aussi être considéré comme un indicateur de taille d'effet*.
- On voit aussi que pour les tableaux 2 x 2, le phi de Cramér se réduit à l'indice défini au point précédent.
- Lorsque les fréquences attendues sont faibles, il peut être bon de regrouper certaines catégories, mais cette opération ne doit pas être motivée par le constat décevant de résultats « non satisfaisants ».

E.1.3. Extension 1 : analyse d'une table de contingences issue de classements d'experts - le « *kappa de Cohen* ».

La « méthode des juges » permet de faire évaluer des objets ou des personnes par un certain nombre de juges. On obtient ainsi soit des *rangs* (que l'on peut comparer par des méthodes non-paramétriques), soit des *classements*, que l'on peut représenter dans une table de contingences. C'est ce dernier cas qui nous intéresse ici, dans le cas particulier de 2 juges ayant classé n objets ou individus. L'analyse du *kappa de Cohen* permet de se faire une idée de *la force de l'accord entre les deux juges*, étant donné que dans une telle situation, un test d'indépendance n'a aucun intérêt (établir l'indépendance des juges reviendrait à dire qu'ils émettent leurs jugements au hasard).

Prenons l'exemple (*cf.* Howell, 1998) d'une classe de 30 adolescents qui se trouvent classés en 3 catégories « cliniques » par 2 juges experts en la matière.

Les trois catégories sont : A (pas de problèmes) ; B (retrait-dépression) et C (agitation-manie) ; on se demande si les avis des deux experts concordent à propos de ce groupe.

Si l'on croise les évaluations, on obtient le tableau suivant :

TABLEAU 8. Répartition de jugements de 2 juges à propos de 30 sujets

	Juge 1			
Juge 2	A	B	C	total
A	15 (10.67)	2	3	20
B	1	3 (1.20)	2	6
C	0	1	3 (1.07)	4
total	16	6	8	30

- Le premier expert classe 16 sujets dans A, 6 dans B et 8 dans C
- Le second expert classe 20 sujets dans A, 6 dans B et 4 dans C.

Les deux juges sont d'accord pour classer 15 élèves dans A, 3 dans B et 3 dans C.

- Par contre, 2 élèves que le juge 2 déclarait A sont en B pour le juge 1.
- et 3 élèves que le juge 2 déclarait dans A sont dans C pour le juge 1.
- de même que 2 élèves classés en B par le juge 1 sont en C pour le juge 2.
- Quant à lui, le juge 1 trouve « sans problèmes » un élève que le juge 2 classe en B
- et le juge 1 classe en B un élève que le juge 2 classe en C.

On constate finalement que les deux juges sont d'accord dans 21 cas sur 30, soit pour 70% des cas. Cette valeur n'est pourtant pas très intéressante car on voit bien que la catégorie « pas de problèmes » recueille une majorité des suffrages. Il faut donc tenir compte des effectifs marginaux et imaginer que les juges pourraient classer au hasard les élèves, tout en respectant la répartition globale entre A, B et C.

- Le juge 1 répartirait au hasard, mais en respectant les proportions de 16 A, 6 B et 8 C.
- Le juge 2 répartirait aussi au hasard, mais en respectant les proportions de 20 A, 6 B et 4 C.

Cette situation reflète une *hypothèse d'indépendance* entre les classements opérés par les deux juges. Comme dans le cas des tableaux de contingences habituels, on peut alors calculer les effectifs attendus en cas d'indépendance (sous H_0).

Par exemple, l'effectif théorique de la case AxA est de (20x16)/30 = 10.67, ce nombre indique le nombre d'accords (à propos de la catégorie A) entre les juges, s'ils avaient classé au hasard tout en respectant les effectifs marginaux.

Il est clair que ce calcul n'a un intérêt que pour les cases situées dans la diagonale du tableau, puisqu'on s'intéresse à un degré d'*accord* (*cf.* Tableau 4). Finalement, on peut constater que le hasard seul classerait : 12.94/30 = .43 = 43% des sujets, ce qui n'est pas négligeable !

Ce qui nous intéresse finalement dans cette affaire, c'est le *degré d'accord entre juges, après correction de l'effet dû au hasard*. On doit à Cohen une formule qui permet de connaître cette valeur, il s'agit d'un indice kappa :

$$\kappa = \frac{ED_o - ED_a}{n - ED_a}$$

où n est l'effectif de l'échantillon, ED_0 est la somme des effectifs diagonaux observés, et ED_a est la somme des effectifs diagonaux « attendus », c'est-à-dire le nombre de concordances dues au seuil hasard de l'échantillonnage.

On se rend compte que par rapport à la proportion que nous avons calculée plus haut, à savoir ED_0 / n, la formule de Cohen corrige ce rapport en soustrayant ED_a au numérateur comme au dénominateur. Le κ vaut ici :

$$\frac{21 - 12.94}{30 - 12.94} = 0,47$$

Il faut être attentif au fait que kappa n'est pas un % d'accord, il mesure en fait un *taux d'amélioration par rapport au hasard*. Son *niveau de signification* dépend du nombre de sujet jugés et n'est que rarement discuté, par contre son *ampleur* doit être interprétée. Certains auteurs (Gendre, 1976) ont donné des appréciations de kappa dans le domaine de la méthode des juges :

- κ compris entre 0 et .20 : est considéré comme *faible* ;
- κ compris entre 0.21 et .40 : est considéré comme *non négligeable* ;
- κ compris entre 0.41 et .60 : est considéré comme *modéré* ;
- κ compris entre 0.61 et .80 : est considéré comme *élevé* ;
- κ compris entre 0.81 et 1 : est considéré comme *exceptionnel* ;

Ces repères doivent toutefois être relativisés selon les domaines dans lesquels ils sont appliqués. En orientation professionnelle, par exemple, les exigences sont inférieures et un indice de .50 est déjà considéré comme exceptionnel.

E.1.4. Extension 2 : analyse d'une table de contingences comportant des effectifs très inégaux - le « *rapport de chances* ».

Appliqué à certaines tables de contingences 2x2 et d'effectifs très inégaux, le « test du chi carré » ne fournit parfois que de maigres informations et il peut être utile d'appliquer d'autres méthodes, très simples et souvent plus fructueuses.

Dans un exemple cité par Howell (1998, p. 182) il est question de l'effet préventif de la prise d'aspirine sur l'occurrence de crises cardiaques chez les hommes. Plus de 22000 médecins se sont prêtés à l'expérience, et la moitié d'entre eux a pris régulièrement une certaine dose d'aspirine, les autres un placebo. Après une certaine période, on a enregistré l'incidence de crises cardiaques et le tableau suivant a pu être dressé :

TABLEAU 9. Incidence de crises cardiaques en fonction de la prise d'aspirine (hommes)

	Crise cardiaque	Pas de crise cardiaque	total
Aspirine	104	10933	11037
Placebo	189	10845	11034
total	293	21778	**22071**

$d^2 = 25$ est significatif au seuil 5%, il y a bien une relation entre la prise d'aspirine et le taux de crises cardiaques, mais comment la caractériser ? Quelle est l'utilité pratique de cette observation ?

On peut utiliser la mesure d'association $\varphi = \sqrt{\frac{25}{22071}} = 0.033$ qui ne donne rien de convaincant avec un effectif si important : la taille d'effet est ridicule.

Par contre, on peut s'intéresser au rapport des « chances » de pas avoir de crise par rapport au fait d'en avoir subi une : c'est le rapport de :

- 10933/104 chez les sujet ayant pris l'aspirine, soit 105.1, et de :
- 10845/189 chez les sujets n'en ayant pas pris, soit 57.38

d'après ces chiffres, on a donc 105.1/57.38 = 1.83 fois plus de chances de ne pas avoir de crise cardiaque en prenant de l'aspirine que si l'on en prend pas... Et voici que ces chiffres prennent soudain un autre sens, très pratique, au point que plusieurs médecins prescrivirent d'office de l'aspirine à tous leurs patients mâles suspectés d'avoir des problèmes vasculaires (et ceci malgré un *phi* dérisoire, rappelons-le).

E.2. Indépendance entre une variables catégorielles et une variable numérique continue

E.2.1. Cas 1 : tests d'indépendance entre une variable numérique et une variable catégorielle dichotomique - *le test de Student.*

Un grand nombre de techniques d'analyse de données s'intéressent aux moyennes d'échantillons dans le but de les comparer, soit à une valeur théorique (*cf.* test d'ajustement déjà abordé ci-dessus), soit à une ou plusieurs autres moyennes d'échantillons.

Les questions qui se posent au sujet des moyennes sont en général de deux ordres :

- Situation 1 : on veut savoir si deux *groupes indépendants* distingués par une caractéristique (sexe, classe d'âge ou toute autre indiquée par une variable dichotomique) varient en moyenne selon une dimension continue (taille, aptitude numérique, trait de personnalité, etc.).
- Situation 2 : on cherche à connaître l'effet d'un « traitement » sur la mesure d'une dimension continue mesurée avant et après dans un même groupe (les deux échantillons sont de ce fait constitués des mêmes individus et sont donc dits « dépendants » ou *appariés*.

- *Situation 1 : deux groupes tirés de manière indépendante*

Comparer deux moyennes ou deux variances d'échantillons ne signifie pas que l'on se demande si elles sont identiques (la probabilité qu'elles le soient rigoureusement est nulle !), mais la vraie question est de savoir si elles ne sont pas trop différentes, ce qui nous permettrait de ne pas rejeter l'hypothèse nulle d'indépendance, à savoir que les deux échantillons proviennent de la même population.

En effet, postuler l'indépendance des deux variables en jeu, celle indiquant l'appartenance au groupe et celle mesurant une dimension continue, revient à dire que la connaissance du groupe d'appartenance ne permet pas de prédire la valeur à la dimension continue – et inversement, la connaissance de la valeur de la dimension continue ne permet pas de deviner l'appartenance à l'un ou l'autre groupe.

Postuler une H_0 d'indépendance revient donc à postuler que les deux échantillons proviennent d'une même population de moyenne μ et de variance σ^2. La comparaison de deux distributions du caractère X mesuré dans deux échantillons suppose donc la comparaison de deux variances *et* de deux moyennes.

Logiquement, le test sur les variances précède celui sur les moyennes, car ce dernier (appelé test du t de Student) n'est pertinent que si les variances des échantillons sont suffisamment proches, c'est-à-dire qu'elles peuvent toutes deux être considérées comme deux estimations de la *même* variance théorique.

- *Principe du test sur les variances :*

- On pose H_o : les deux variances empiriques S_1^2 et S_2^2 (de deux échantillons de taille n_1 et n_2) sont deux estimations de la même variance théorique σ^2 ;

- autrement dit, H_o s'écrit : $\sigma_1^2 = \sigma_2^2 = \sigma^2$; (homogénéité des variances)

- Si H_o est vraie, le rapport S_1^2 / S_2^2 (on met la plus grande variance au numérateur) suit une loi de F (Fisher-Snedecor) à $(n_1 - 1)$ et $(n_2 - 1)$ degrés de liberté;

- Si le rapport S_1^2 / S_2^2 dépasse un seuil $f_{(1-\alpha)}$ fixé, l'H_o est rejetée avec un risque d'erreur α.

- *Principe du test sur les moyennes (« test du t de Student ») :*

- On pose H_o : les deux moyennes empiriques M_1 et M_2 (de deux échantillons de taille n_1 et n_2) sont deux estimations de la même moyenne théorique μ ;

- autrement dit, H_o s'écrit : $\mu_1 = \mu_2 = \mu$;

- Si H_o est vraie, M_1 est une variable échantillonnale d'espérance μ et de variance : $\frac{S_1^2}{n_1}$. Comme déjà vu dans le cas du test d'ajustement, cette variable suit une loi de t (Student), et non une loi gaussienne car la variance de la population est estimée par celle de l'échantillon ;

- de même pour M_2, dont la variance est : $\frac{S_2^2}{n_2}$;

- Toujours si H_o est vraie, la variable $M_1 - M_2$ suit aussi une loi t, d'espérance zéro et de variance égale à la somme des variances des échantillons[19] : $S^2_{M_1 - M_2} = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$:

- Mais si H_o est vraie, les deux variances empiriques estiment la même variance théorique σ^2 ; on peut donc calculer une estimation $\hat{S}^2$ (un analogue du $\hat{p}$ dans le

19. La variance de la somme ou de la différence de deux variables *indépendantes* est égale à la somme des deux variances (*cf.* note 12).

cas du test sur les proportions) de σ^2 en pondérant les variances empiriques S_1 et S_2 par leurs degrés de liberté (ou autrement dit par les effectifs des groupes diminués de 1),

cette estimation vaut : $$\hat{S}^2 = \frac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{(n_1 - 1) + (n_2 - 1)}$$

- Donc, la variance de la différence des deux moyennes empiriques vaut :

$$\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} = \hat{S}^2 \cdot \left[\frac{1}{n_1} + \frac{1}{n_2}\right] = \frac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{(n_1 - 1) + (n_2 - 1)} \cdot \left[\frac{1}{n_1} + \frac{1}{n_2}\right]$$

- Si H_o est vraie, la variable $M_1 - M_2$ suit donc une loi t d'espérance zéro et d'écart-type : $$\sqrt{\frac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{(n_1 - 1) + (n_2 - 1)} \cdot \left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$$

- Les tables décrivent les fractiles de la loi de t *standardisée*, il faut donc *centrer* et *réduire* notre variable $M_1 - M_2$ de manière à pouvoir y situer l'une de nos *réalisations* $(m_1 - m_2)$ pour une expérience donnée ;

- Comme la variable $M_1 - M_2$ a une espérance (ou une moyenne) de 0 si H_o est vraie, elle est déjà centrée ; il faut donc encore la réduire en la divisant par son écart-type et finalement, la quantité :

$$T = \frac{M_1 - M_2}{\sqrt{\frac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{(n_1 - 1) + (n_2 - 1)} \cdot \left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}$$

suit une loi de t de Student à $(n_1 - 1) + (n_2 - 1)$ degrés de liberté, d'espérance nulle et d'écart-type n/n-2 (où $n = n_1 + n_2$).

- ... et si une réalisation t de T, pour une expérience particulière, dépasse un seuil $t_{(1-\alpha)}$ (test bilatéral) fixé, alors H_o est rejetée avec un risque d'erreur α.

En résumé : Comparaison de moyennes dans le cas de groupes indépendants

- *Conditions d'utilisation* : Les individus des deux groupes sont tirés au hasard, le caractère X est supposé normalement distribué dans les deux échantillons.

- *Logique des techniques de comparaison de moyennes et de variances* : le problème de fond est de savoir si deux échantillons proviennent d'une même population.

L'hypothèse nulle peut donc porter sur la seule moyenne, ou être plus exigeante et porter sur la moyenne *et* la variance. Comme le test de Student ne fonctionne bien que si les variances sont proches, il vaut mieux toujours commencer par le test des variances.

- Si l'on ne rejette pas l'hypothèse nulle sur les variances, la situation est favorable – et si l'hypothèse nulle sur les moyennes n'est pas non plus rejetée, on a une bonne conviction en faveur de l'unicité de la population d'origine.
- Si l'hypothèse nulle sur les variance doit être rejetée, on n'obtiendra qu'une conviction partielle au sujet de l'origine des échantillons : une différence de moyennes non significative suggère la conclusion que les échantillons proviennent d'une population unique, mais seulement du point de vue de la moyenne, ce qui n'est pas toujours suffisant.
- D'autre part, si les variances diffèrent significativement, les formules du test de Student doivent être adaptées (*cf.* Howell, p. 225).

- *Procédure à suivre* : soient 2 groupes-échantillons de sujets de taille n_1 et n_2 mesurés selon une caractéristique X.
  - Calculer les moyennes et variances empiriques de chaque échantillon ;
  - tester une H_o sur les variances, si elle est rejetée, s'interroger sérieusement sur l'opportunité de continuer à comparer les deux échantillons.
  - Il faut noter que le test sur les variances décrit ci-dessus est *très* sensible à la violation de l'hypothèse de normalité des distributions originales, il convient de lui préférer le *test de Levene*, figurant dans la sortie standard SPSS.
  - Si les différences de variances ne sont pas significatives, calculer l'estimation de la variance théorique en pondérant les variances empiriques par les effectifs :

$$s^2 = \frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

  - calculer ensuite l'écart-type de la différence des moyennes, puis la valeur :

$$t = \frac{m_1 - m_2}{\sqrt{\frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{(n_1 - 1) + (n_2 - 1)} \cdot \left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}$$

  - Si H_o est vraie, t suit une loi t de Student à $n_1 + n_2 - 2$ degrés de liberté, il suffit donc de comparer cette valeur au seuil déterminé par un domaine de rejet de 5% (ou 1%) dans la distribution de t.

- *Remarques* : une fois l'homogénéité des variances établie, la violation des hypothèses de normalité n'a pas grand effet sur les résultats du test sur les moyennes (on dit que le test de Student est *robuste*).

- Plus les échantillons sont grands, plus les corrections à apporter en raison de l'hétérogénéité des variances sont inutiles.
- Pour évaluer la force du lien entre les deux variables dichotomique et continue, on peut déduire un *coefficient de corrélation point bisérial* à partir de t, à l'aide de la formule : $r^2 = \frac{t^2}{t^2 + n_1 + n_2 - 2}$
- Une bonne alternative au test d'hypothèse classique (de plus en plus critiqués dans la littérature scientifique) consiste à calculer un *intervalle de confiance* autour de la différence des moyennes observées, ayant 95 ou 99% de chances de contenir la valeur attendue zéro, si H_0 est vraie. Cet intervalle est centré en $m_1 - m_2$ et sa demi-largeur est égale au produit de l'écart-type de la variable $M_1 - M_2$ par la valeur seuil pour t (95% ou 99%), avec n – 2 degrés de liberté.
- La taille d'effet (dû à l'appartenance aux deux groupes) peut se mesurer de deux manières : soit on utilise la formule de Cohen pour calculer d en utilisant comme estimation de la variance de la population la quantité :

  $$s^2 = \frac{(n_1 - 1) \cdot s_1^{\,2} + (n_2 - 1) \cdot s_2^{\,2}}{(n_1 - 1) + (n_2 - 1)}$$

  ; soit on juge la corrélation *point bisériale* comme un coefficient de corrélation normal..

- *Situation 2 : deux groupes tirés de manière non-indépendante (groupes appariés, situation test-retest)*

On rencontre cette situation lorsque on désire tester l'effet d'un traitement sur un groupe d'individus. La procédure habituelle consiste à mesurer un caractère *avant* un traitement donné, puis à mesurer ce même caractère *après* le traitement, de manière à savoir s'il y a eu évolution de la situation, dans un sens ou dans l'autre.

C'est ainsi que l'on peut mesurer l'effet d'un médicament, d'une formation, d'une thérapie ou de n'importe quel traitement sur les sujet d'un groupe appelé précisément « groupe traitement ». Dans les plans d'expérience classiques, cette comparaison s'effectue parallèlement à l'étude d'un groupe auquel est administré un placebo, ou ne bénéficiant d'aucun traitement, appelé « groupe contrôle ».

Si l'on s'intéresse spécifiquement à l'effet du traitement sur l'un des groupes, on peut procéder à une comparaison de moyennes et se demander si les résultats « avant » se distinguent *significativement* de ceux « après ».

L'hypothèse nulle postule l'inefficacité du traitement : les résultats du groupe sont les mêmes, aux aléas de l'échantillonnage près, avant et après traitement, autrement dit les deux distributions empiriques ne peuvent pas être distinguées, aux aléas de l'échantillonnage près.

- Traitement du problème :

On pourrait croire que les formules développées plus haut vont s'appliquer dans ce cas comme dans les précédents, il n'en est pourtant rien à cause de la dépendance entre les groupes : les variances « avant » et « après » ne sont pas indépendantes puisque calculées sur les résultats des *mêmes* individus mesurés deux fois.

Il s'en suit que l'estimation de la variance de la différence des moyennes ne peut plus se faire simplement car les variances ne sont plus additives, et

on ne peut donc plus écrire que : $S^2_{M_1 - M_2} = \frac{S_1^{\ 2}}{n_1} + \frac{S_2^{\ 2}}{n_2}$

Lorsque deux échantillons ne sont pas tirés de manière indépendante, la variance de la somme (ou d'une différence) de deux variables n'est pas égale à la simple somme des variances !

- *Solution* : on peut contourner la difficulté en ne s'intéressant plus à *la différence des moyennes*, mais en calculant la *moyenne des différences*, sujet par sujet.

- Soit un caractère X mesuré aux temps t_o et t_1, séparés par un traitement.

- L'hypothèse nulle d'inefficacité des traitements se formalise par $H_o : \mu_d = 0$; autrement dit X et X' sont une seule et même variable, ou encore : on ne peut pas, pour un individu donné, distinguer des scores de X « avant » et « après » le traitement.

Sujet	X (à t_o)	X' (à t_1)	$X_{to} - X'_{t1} = d_i$
1	x_1	x'_1	$x_1 - x'_1$
2	x_2	x'_2	$x_2 - x'_2$
3	x_3	x'_3	$x_3 - x'_3$
	etc. jusqu'à…		
n	x_n	x'_n	$x_n - x'_n$

- L'hypothèse nulle d'inefficacité des traitements se formalise par $H_o : \mu_d = 0$; autrement dit X et X' sont une seule et même variable, ou encore : on ne peut pas, pour un individu donné, distinguer des scores de X « avant » et « après » le traitement.

- La *moyenne* des différences observées est une variable M_d ; si H_o est vraie, M_d est normale, a une espérance zéro et une variance : $\frac{\sigma^2}{n}$, mais on ne connaît pas la variance théorique des différences, on va donc l'estimer par la variance des différences observées dans l'échantillon, à savoir : $S_d^{\ 2}$

- et on sait que M_d suit donc une loi de Student
  d'espérance zéro et de variance : $\frac{S_d^2}{n}$

- Finalement, la quantité standardisée : $T = \frac{M_d}{\sqrt{\frac{S_d^2}{n}}}$ suit une loi
  de Student à n – 1 degrés de liberté.

... et si une réalisation t de T, pour une expérience particulière, dépasse un seuil $t_{(1-\alpha)}$ fixé (test unilatéral), alors H_o est rejetée avec un risque d'erreur α

.

En résumé : Comparaison de moyennes dans le cas de groupes non indépendants (situation test-retest ou sujets appariés : frères et soeurs, personnes déclarées semblables selon un critère, etc.)

- *Conditions d'utilisation* : pas de condition particulière hormis la dépendance entre sujets des groupes.
- *Procédure à suivre* : soient n sujets mesurés deux fois à propos d'un caractère X. X est la première mesure, X' la seconde.
  - Pour chacun des sujets, calculer la différence $d_i = x_i - x'_i$
  - Puis calculer la moyenne des d_i = $m_d = \frac{1}{n} \cdot \sum_1^n (x_i - x'_i)$
  - Et la variance des d_i = $s_d^2 = \frac{1}{n-1} \cdot \sum_1^n (d_i - m_d)^2$
  - Calculer ensuite la quantité : $t = \frac{m_d}{\frac{s_d}{\sqrt{n}}}$
  - Si H_o est vraie, t suit une loi de t de Student à n-1 degrés de liberté, il suffit donc de comparer cette valeur au seuil déterminé par un domaine de rejet de 5% (ou 1%) dans la distribution de t de la table, correspondant au nombre de degrés de liberté n – 1.
  - *Remarque* : on veillera à bien déterminer les seuils en fonction des tests uni - ou bilatéraux, selon qu'on attend X' plus grand que X, ou inversement, ou que l'on a pas d'attente particulière

- On peut calculer une estimation acceptable de la taille de l'effet en divisant la différence des moyennes des deux variables par l'écart-type de toutes les observations.
- Howell donne une formule plus précise : $d = \dfrac{m_1 - m_2}{s \cdot \sqrt{(1-\rho) \cdot 2}}$; où sont les moyennes des deux passations (s'il s'agit d'un test), s est la variance de l'une des passations et r la corrélation entre les deux, donc la fidélité du test.

Exemple :

Voici les données d'un groupe de 5 sujets testés deux fois sur une caractéristique X, une fois avant (à t_o) un certain traitement, et une fois après (à t_1). On se demande si le traitement a eu un « effet » positif.

Comme on attend une augmentation de X en moyenne, on postule deux hypothèses complémentaires :

H_o : le traitement est sans effet, donc $\mu_d = 0$; contre une alternative :
H_1 : le traitement a un effet positif, donc $\mu_d > 0$

Tableau des observations :

Sujet	X (à t_o)	X' (à t_1)	$X_{to} - X'_{t1} = d_i$
1	3	4	1
2	2	4	2
3	5	7	2
4	6	8	2
5	2	5	3

La moyenne des différences vaut : (1+2+2+2+3) / 5 = 2

La variance des différences vaut :
$[(1\text{-}2)^2 + (2\text{-}2)^2 + (2\text{-}2)^2 + (2\text{-}2)^2 + (3\text{-}2)^2] / 4 = [1+0+0+0+1] / 4 = 1/2.$

La variable de décision M_d a donc une espérance de 0

et (pour notre expérience) un écart-type de $\sqrt{\dfrac{1}{2 \cdot 5}} = \sqrt{\dfrac{1}{10}}$

La valeur de t standardisée vaut donc ; $t = \dfrac{2}{\sqrt{\dfrac{1}{10}}} = 2 \cdot \sqrt{10} = 6,32$

cette valeur doit être comparée à la valeur de t à 4 degrés de liberté au seuil (unilatéral à droite) 5%. On regarde donc la table dans la colonne 10% (!!!) et on voit que le seuil 5% unilatéral à droite est 2.13. H_o est donc rejetée au profit de H_1 : le traitement semble efficace.

On peut aussi calculer un intervalle de confiance à 95% autour de m_d, réalisation de M_d dont l'espérance est zéro et l'écart-type = $\sqrt{\frac{1}{10}}$ = 0,316 ;

l'intervalle de confiance se détermine donc comme suit :

$$[2-(t_{0.95}\cdot 0,316);2+(t_{0.95}\cdot 0,316)] = [2-(0,67);2+(0,67)] = [1,33;2,67]$$

… et on voit bien qu'il ne contient pas la valeur zéro.

E.2.2. Cas 2 : tests d'indépendance entre une variable numérique et une variable catégorielle quelconque (plusieurs niveaux) - *le test du F de Fisher-Snedecor et l'« analyse de variance »*

On regroupe sous le terme « analyse de variance » une grande diversité de techniques qui ont toutes pour but de distinguer si deux, trois ou plus de trois groupes peuvent être considérés comme ayant été tirés d'une seule et même population.

S'il n'y a que deux groupes, on se retrouve dans le cas de la comparaison de deux moyennes par le biais d'un test de Student, mais l'analyse de la variance peut aussi être appliquée.

C'est principalement lorsque l'on est en présence de trois ou davantage de groupes que l'analyse de variance s'impose : il est en effet très peu judicieux de distinguer plusieurs groupes au moyen de tests de t successifs : les groupes n'étant pas indépendants, les tests sont liés et les niveaux de signification des divers tests se contaminent les uns les autres, si bien qu'il faudrait leur apporter des corrections qui ne sont pas toujours simples.

Lorsqu'on est en présence de plusieurs groupes, ceux-ci peuvent être déterminés par une seule variable catégorielle (appelée « facteur »), ou par plusieurs facteurs dont les niveaux se croisent.

- Par exemple : une variable catégorielle à trois niveaux (bas/moyen/élevé) détermine évidemment trois groupes. Mais si on la croise avec une variable dichotomique (par ex : sexe F ou M), on est en présence de 6 groupes (« F-bas »/« M-bas »/« F-moyen »/etc.). Si l'on ajoute encore l'influence d'un troisième facteur (âge : <15 ans/>15 ans), on se retrouve avec 12 groupes… Et les choses deviennent rapidement très complexes.

- Les expériences qui spécifient soigneusement les combinaisons de variables et qui s'intéressent à la moyenne d'une variable continue dans chacun des groupes doivent décrire un *plan factoriel* précis, dans lequel il doit être spécifié si les niveaux des facteurs sont fixes ou aléatoires, si les mesures sont successives (appariées) ou non, si un facteur est « niché » dans un autre ou non, et si on s'intéresse aussi à l'effet des facteurs sur une deuxième, voire une troisième variable continue (appelée alors « covariable ») auquel cas on doit s'aventurer du côté des techniques d'analyse de variance *multiples* (MANOVA) qui ne sont pas toujours simples à saisir pour des non-statisticiens.

- En bref, un tour d'horizon des diverses techniques de l'analyse de variance nécessiterait un cours complet à lui seul et nombreux sont les ouvrages volumineux qui y sont consacrés. À dire vrai, l'analyse de variance est surtout pratiquée par les expérimentalistes qui en ont fait une sorte de « religion » caractérisée par des rites et des terminologies parfois différentes, ce qui ne facilite pas son abord par les non-initiés…

- Dans ce cours nous nous contenterons d'expliciter le principe fondamental de l'analyse de variance, et cela dans le cadre le plus simple : celui de l'analyse de variance simple à un seul facteur de classification et une seule variable dépendante.

- Il n'est peut-être pas inutile de rappeler que l'analyse de variance est une technique de décomposition de la variance des scores individuels qui a pour but de déterminer si des groupes diffèrent selon leurs *moyennes* ! Ce test devrait effectivement être accompagné d'un test sur les variances, comme dans le cas du test de Student, car si plusieurs groupes sont censés être tirés de la même population, on attend au moins qu'ils aient même moyenne et même variance. Pour simplifier les choses, le test de Levene, qui porte sur l'homogénéité des variances s'effectue par le bais d'un test sur les moyennes (!).

- *Plan simple : un seul facteur de classification*

Il est important d'expliciter préalablement les hypothèses de base de l'analyse de variance :

- L'hypothèse nulle du test sur les moyennes veut que les groupes soient tirés de la même population, autrement dit, que les moyennes observées pour chacun des n groupes soient les estimations de la même moyenne μ dans la population.
  (On écrit couramment que $H_o : \mu_1 = \mu_2 = \mu_{i = \dots =} \mu_{n =} \mu$)

- Le test d'hypothèse standard sur les moyennes n'est en fait valable que si les différents groupes ont même effectif et même variance. En général *on s'arrange* pour que les effectifs soient à peu près égaux et *on présuppose* que les variances des groupes sont à peu près égales.

- Il faut savoir qu'en psychologie, il est parfois difficile de satisfaire à ces deux conditions, on calmera toutefois nos scrupules en utilisant des logiciels qui tiennent

compte des différences d'effectifs, d'une part, et en testant l'homogénéité des variances empiriques, d'autre part. Cela dit. lorsque le test de Levene aboutit au rejet de l'hypothèse d'homogénéité des variances, on se gardera de tirer des conclusions trop définitives de notre expérimentation.

- Une dernière condition d'utilisation porte sur la normalité des distributions du caractère numérique. À vrai dire, peu d'utilisateurs s'en préoccupent vraiment, prétextant que l'analyse de variance est une technique « robuste ».

Il convient maintenant de distinguer *cinq étapes fondamentales* dans le processus d'« analyse de la variance ».

1. La première étape est purement *descriptive* : il s'agit avant tout d'examiner les moyennes empiriques (m_j) des groupes et les comparer à la moyenne générale (m_T), on peut ainsi se faire une première idée de l'effet du facteur sur la variable numérique, et identifier immédiatement le ou les groupes susceptibles de se distinguer des autres.

2. La deuxième étape est *analytique,* elle consiste à décomposer *l'information*, c'est-à-dire l'écart entre le score de chaque individu (x_{ij}) et la moyenne générale (= M_T, une variable, si l'on raisonne en toute abstraction, avant toute réalisation).

On reconnaîtra sans peine que l'écart total (d'un score quelconque à la moyenne générale) peut se décomposer en un écart « intra-groupe » (du score à la moyenne de son groupe) + un écart « inter-groupe » (de la moyenne du groupe à la moyenne générale. Si l'on divise la somme des carrés de ces écarts (**S**um of **S**quares…) par N, nombre de sujets, on est en présence des trois variances : totale, intra– et inter-groupes, qui sont additives.

- La *variance totale* = **SST/N** (**S**um of **S**quares **T**otal/N) est la variance des écarts des scores individuels à la moyenne générale.
- La *variance intragroupe* = **SSW/N** (**S**um of **S**quares **W**ithin/N) est la variance des écarts des scores à la moyenne de leur groupe, elle est considérée comme de l'erreur et il faut la considérer comme la variance échantillonnale habituelle des scores dans chaque groupe.
- La *variance intergroupe* = **SSB/N** (**S**um of **S**quares **B**etween/N) est la variance due à l'appartenance aux groupes, c'est la variance des écarts des moyennes des groupes à la moyenne générale. On peut aussi dire que c'est la part de variance due à l'*effet* du facteur sur la variable numérique.

On peut alors écrire l'équation de l'analyse de la variance qui permet de comprendre l'information totale apportée par la variable continue comme étant égale à la somme d'une information « explicable » (variance intergroupes) et d'une part de variance d'erreur (variance intragroupe). Chaque score individuel peut donc aussi être décomposé en sa partie « explicable » et sa partie d'« erreur », et par conséquent, les quatre expressions suivantes sont strictement équivalentes :

1. $Xij = (X_{ij} - M_j) + (M_j - M_T) + M_T$

2. *SST/N (variance tot) = SSW/N (variance intra) + SSB/N (variance inter)*

3. $$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-M_T)^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-M_j)^2 + \sum_{j=1}^{k} n_j \cdot (M_j-M_T)^2$$

4. *SST = SSW + SSB*

3. La troisième étape a des *finalités pratiques*, elle consiste à évaluer la force du lien existant entre le facteur et la variable numérique, autrement dit à mesurer l'intensité de l'« effet » du facteur. Cette mesure peut s'effectuer par l'intermédiaire d'un coefficient « éta » qui est l'analogue d'un coefficient de corrélation. Sachant que le carré d'une corrélation mesure le % de variance totale d'une variable « expliquée » par l'autre, on peut calculer, dans le contexte de l'analyse de variance, la part de variance totale due à la variance « explicable », c'est à dire :

$$\eta^2 = \frac{\text{Var(INTER)}}{\text{Var(TOTALE)}} = \frac{\text{Var(INTER)}}{\text{Var(INTER)} + \text{Var(INTRA)}} = \frac{SSB}{SSB + SSW}$$

Comme dans le contexte de la régression, éta carré exprime un « % de variance expliquée », et éta s'interprète comme un coefficient de corrélation usuel *ce qui en fait une bonne estimation de la taille de l'effet dû à l'appartenance aux groupes*.

4. La quatrième étape est *inférentielle*, car il reste maintenant à savoir si cet effet, mesuré par « éta », peut être réellement attribué à l'effet du facteur dans la population, ou s'il est simplement dû aux aléas d'échantillonnage.

Cette question revient à se demander si les moyennes empiriques ne varient qu'en raison de l'échantillonnage, autrement dit, si les groupes sont tirés de la même population. Cette hypothèse est l'H_0 du test de F associé à l'analyse de variance.

Le principe du *test de F* est le suivant :

- On peut admettre ou supposer, sans exprimer une hypothèse particulière, que les variances empiriques des k groupes (d'effectif total N) sont toutes des estimations de la variance σ^2 de la population. Donc leur moyenne pondérée (en fonction des effectifs) est aussi une estimation de cette variance. Or cette moyenne pondérée vaut :

$$\frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-M_j)^2}{N-k} = \frac{SSW}{N-k}$$

on l'appelle « Mean SSW » (MSSW) et elle ne doit pas être confondue avec la variance intragroupes (qui vaut SSW/N).

- Toujours si H_o est vraie, et *en vertu du théorème central limite*, la variance des moyennes estime σ^2/n , donc :

$$\frac{\sum_{j=1}^{k} n_j \cdot (M_j - M_T)^2}{k-1} = \frac{SSB}{k-1}$$

estime aussi σ^2, cette quantité est appelée Mean SSB (MSSB), et elle ne doit pas être confondue avec la variance intergroupes (qui vaut SSB/N).

- donc MSSW et MSSB estiment toutes deux σ^2. Afin de ne pas confondre les « Mean Squares » avec les variances intra et inter, on notera que les Mean Squares ne sont pas additives !
- Nous sommes donc en présence de deux estimations de la même variance théorique, or nous savons que le rapport de ces deux estimations (la plus grande, MSSB, étant au numérateur) suit une loi de F avec [df de MSSB ; df de MSSW] degrés de liberté.
- Il ne reste alors qu'à réaliser une expérience, calculer SSB et SSW, puis éta, puis calculer les « Mean Squares »[20] et former le quotient F. On lit ensuite dans la table du F de Fisher afin de vérifier que la valeur de F = MSSB/MSSW ne dépasse pas un seuil convenu.
- Si tel était le cas, c'est-à-dire si F empirique dépasse une valeur $F_{[(k-1);(n-k)]\ (1-0.05)}$, alors H_o peut être rejetée au seuil 5% : éta mesure un lien non nul, les groupes ne sont pas homogènes du point de vue de la variable dépendante et ne sont donc pas tous tirés de la même population.

5. La dernière étape de l'analyse consiste à décrire les différences en vue de les interpréter. Il s'agit de savoir lequel (ou lesquels) des groupes se distinguent « significativement » des autres. Cette question peut être résolue par l'intermédiaire de comparaisons « post-hoc » qui s'effectuent au moyen de *tests de Scheffé* (entre autres). Ces techniques comparent les groupes deux à deux tout en ajustant le niveau de signification des tests.

- *Plans factoriels complexes : plusieurs facteurs*

Nous n'aborderons ici que très brièvement les plans permettant d'analyser l'effet de deux facteurs sur une variable dépendante. Il est clair que si le facteur A comporte k niveaux, et le facteur B n niveaux, le plan factoriel comportera k x n cellules dont il est possible de calculer la moyenne.

20. Les logiciels courants ne calculent pas les variances, mais seulement les « sum of squares » et les « mean square ».

L'analyse de la variance des scores individuels consiste à « décortiquer » les écarts de chaque score à la moyenne générale en :

- un écart dû à un éventuel effet du facteur A,
- un écart dû à un éventuel effet du facteur B,
- un écart dû à un éventuel effet d'interaction entre les niveaux de A et de B,
- un écart « résiduel » dû à la présence incontournable d'un aléa d'échantillonnage (variance intragroupe).

Chacun de ces effets peut être évalué au moyen d'un coefficient éta et d'un test de signification approprié.

Tout le problème consiste alors à savoir se servir d'un logiciel statistique et de connaître suffisamment bien le jargon de l'analyse de variance pour comprendre et interpréter les sorties.

Exemple :

Afin de bien saisir ce que l'on entend par « analyse des effets » on peut prendre un exemple relativement simple :

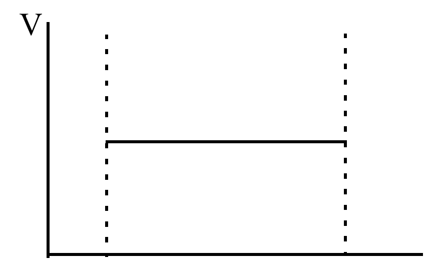
- Soit un facteur A (*traitement = groupe)* à trois niveaux :
  médicament (1)
  placebo (2)
  contrôle (3)

- et un facteur B (*test = occasions*) à deux niveaux :
  avant traitement (1)
  après (2)

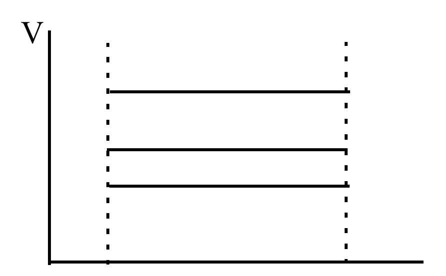
La variable dépendante V étant le *niveau d'anxiété*, par exemple.

On peut observer huit cas de figure représentés graphiquement ci-dessous :

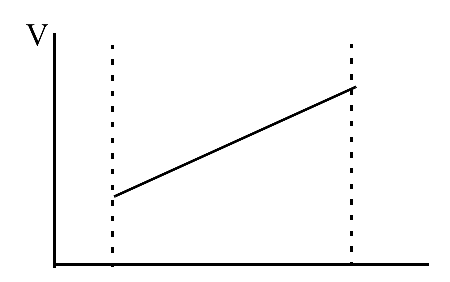
1. Pas d'effet de A, pas d'effet de B, pas d'interaction :



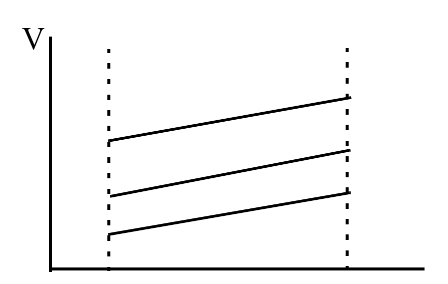
2. Effet de A, pas d'effet de B, pas d'interaction :



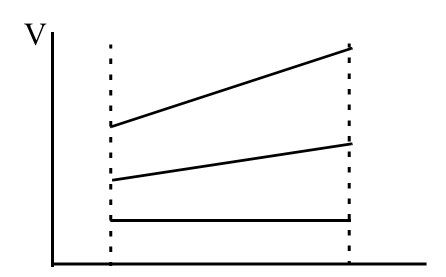
3. Pas d'effet de A, effet de B, pas d'interaction :



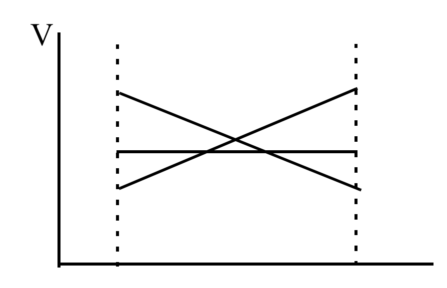
4. Effet de A, effet de B, pas d'interaction :



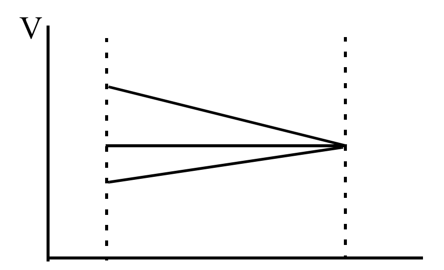
5. Effet de A, effet de B, interaction :



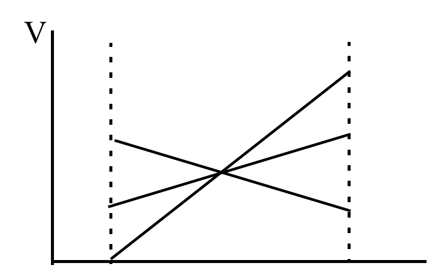
6. Pas d'effet de A, pas d'effet de B, mais interaction :



7. Effet de A, pas d'effet de B, interaction :



8. Pas d'effet de A, effet de B, interaction :



E.3. Indépendance entre deux variables numériques continues

Le coefficient de corrélation utilisé comme statistique d'un *test d'ajustement à l'hypothèse d'indépendance.*

Soit une série de n paires d'observations effectuées sur n sujet ou objets tirés au hasard dans une population. Les observations se rangent en deux variables X et Y, si possible normalement distribuées. Soient m_X et m_Y les moyennes et s_X et s_Y les écart-types empiriques de ces deux variables.

- On appelle *covariance empirique* de X et de Y la quantité :

$$\text{cov}_{(XY)} = \frac{1}{n} \cdot \sum_{1}^{n} (X_i - M_X) \cdot (Y_i - M_Y)$$

- La *corrélation* entre X et Y est une mesure standardisée de la force du lien existant entre les deux variables. La corrélation (cor(X;Y) ou $r_{(X;Y)}$ ou simplement *r*) varie entre –1 et +1, et n'est rien d'autre que la covariance de X et Y *standardisés* :

$$\text{cov}(z_X;z_Y) = \frac{1}{n} \cdot \sum_{i=1}^{n} (z_{Xi} - 0) \cdot (z_{Yi} - 0) = \frac{1}{n} \cdot \sum_{i=1}^{n} z_{Xi} \cdot z_{Yi} \qquad \text{donc :}$$

$$\mathrm{cov}(z_X;z_Y) = \frac{1}{n}\cdot\sum_{i=1}^{n} z_{Xi}\cdot z_{Yi} = \frac{1}{n}\cdot\sum_{i=1}^{n}\left(\frac{X_i - m_X}{s_X}\right)\cdot\left(\frac{Y_i - m_Y}{s_Y}\right) \quad \text{et}$$

$$\mathrm{cov}(z_X;z_Y) = \frac{1}{n}\cdot\sum_{i=1}^{n}\left(\frac{(X_i - m_X)\cdot(Y_i - m_Y)}{s_X\cdot s_Y}\right) = \frac{\mathrm{cov}(X;Y)}{s_X\cdot s_Y} = r_{(XY)}$$

- Les tests d'ajustement à des coefficients de corrélation théoriques non nuls exigent une transformation préalable de r (*cf.* Howell p. 292) et ne seront pas traités ici. Nous nous contenterons de présenter l'ajustement le plus simple et le plus courant, c'est-à-dire celui d'une valeur r observée, à la valeur théorique *zéro*. Ce type d'ajustement revient donc à *tester l'indépendance* de X et de Y.

- On admettra sans démonstration que si H_0 est vraie (indépendance de X et de Y), alors r calculé sur un échantillon suit une loi à peu près normale et d'espérance zéro. Plus précisément, Fisher a montré que la quantité :

$$t = \frac{r\cdot\sqrt{n-2}}{\sqrt{1-r^2}}$$

suit une loi de Student à n – 2 degrés de liberté. Il faut noter qu'il existe des tables spécialement conçues pour lire le niveau de signification d'un coefficient de corrélation (*cf.* table : valeurs critiques pour r…).

Exemple :

Dans un échantillon de 27 individus tirés au hasard dans une population déterminée, on mesure une corrélation de .58 entre deux tests d'aptitudes. Peut-on considérer que chacun de ces tests valide l'autre ?

- On pose H_0 : les résultats aux deux tests sont indépendants, la corrélation entre les variables est nulle dans la population.

- On calcule ensuite :

$$t = \frac{0,58\cdot\sqrt{27-2}}{\sqrt{1-0,58^2}} = 3.56$$

- La valeur 3.56 dépasse le seuil fixé pour $\alpha = 5\%$, à savoir 2.06 lu dans la table à la ligne correspondant aux degrés de liberté 25.

- La corrélation est alors déclarée « significative » ce qui signifie que l'on peut rejeter l'hypothèse nulle avec un risque (de première espèce) égal à 5%.

- *Plus simplement*, on peut aussi regarder dans une table la limite inférieure d'une corrélation significative au seuil 5% mesurée dans un échantillon de taille 27. On

peut y lire (colonne 0.05 et ligne 27–2 = 25) que cette valeur limite est .38. Comme notre valeur empirique (.58) est supérieure, l'hypothèse nulle d'indépendance peut être rejetée.

F. De la dépendance linéaire à la « prédiction »

F.1. Cas 1 : le coefficient de corrélation utilisé comme paramètre d'un modèle prédictif - *modèles de régression simple*

- *Modèles prédictifs : définition*

Les modèles de régression linéaire (ou non linéaire) sont des équations algébriques (du premier degré) permettant de combiner de manière optimale une ou plusieurs variables (dites : prédictrices ou indépendantes) de manière à approximer au mieux une variable critère (ou dépendante).

Lorsqu'il y a plusieurs prédicteurs, ces équations algébriques représentent des *combinaisons linéaires*. Il existe également des modèles de *régression non linéaires* (modèles logistiques, quadratiques, exponentiels, etc.), mais ceux-ci sont encore relativement peu utilisés en sciences sociales, et particulièrement peu en psychologie[21].

Le terme de modèle de régression peut induire en erreur, à vrai dire il est impropre et représente le résidu terminologique d'une théorie fausse, due à Galton (théorie de la « régression vers la moyenne »). En fait on devrait utiliser le terme plus approprié de *modèles de prédiction*. Donc :

les modèles de prédiction cherchent à modéliser de manière optimale les liens entre une ou plusieurs variables considérées comme « prédictrices », et une variable « à prédire » appelée critère.

Deux situations peuvent se présenter : soit on étudie le lien entre le critère et *une* variable prédictrice (prédiction simple), soit on s'intéresse aux liens entre le critère et *plusieurs* variables prédictrices. Il va de soi que c'est le chercheur seul qui décide d'attribuer les rôles de prédicteur ou de critère. N'importe quelle variable numérique ou ordinale peut jouer ces deux rôles, tout dépend des objectifs poursuivis, qu'ils soient d'ordre pratique ou théorique. Notons encore que si la variable critère est qualitative, on n'utilisera plus des modèles de régression, mais des *modèles discriminants* qui sont en général étudiés dans le cadre des théories et techniques dites de *classification* (par contraste avec le terme de *régression)*.

21. Une exception notable : la théorie des tests basés sur des modèles stochastiques, tels le modèle de Rash, Birnbaum, etc.

- *Modèles de régression linéaire simple*

Supposons que l'on s'intéresse au lien existant entre un prédicteur P (variable indépendante), par exemple une note à un test – et un critère C (variable dépendante), par exemple une note d'examen.

Construire un modèle de prédiction de C par P consiste à trouver une équation linéaire en P, permettant d'approcher C, au mieux.

Supposons que cette équation existe. Comme elle est linéaire, elle est de la forme : $a \cdot P + b = \hat{C}$ où a est un nombre, appelé *coefficient de régression*, P est une *variable* et b est une constante appelée *intercept*. Il va de soi que cette relation peut être représentée graphiquement sous forme d'une droite dont a est la *pente* et b l'*ordonnée à l'origine*.

La *contrainte d'ajustement* que nous nous imposons implique que $\hat{C}$ soit aussi « proche » de C que possible. En statistique, la proximité de deux variables s'assimile à la force de leur lien et se représente par le coefficient de corrélation r calculé entre $\hat{C}$ et C. La différence entre $\hat{C}$ et C, que l'on espère minimum pour chaque cas s'appelle le *résidu* ou l'*erreur*.

Pour expliciter l'équation de prédiction, il nous faut donc trouver les coéfficents a et b de l'équation ci-dessus, *de telle manière à ce que la corrélation de* $\hat{C}$ *et* C *soit maximum*. La théorie statistique permet d'*estimer* a et b sur la base d'un échantillon, ce qui permet, à certaines conditions (tirage aléatoire, normalité des distributions de C et P, etc.) d'en inférer que le modèle (équation) de régression estimé sur la base de cet *échantillon d'apprentissage* est utilisable pour tout nouveau cas pris dans la population.

- *Par exemple* : on demande à un échantillon d'élèves, en principe choisis aléatoirement, de passer un test (P) et on note le résultat d'un examen ultérieur (C). Cet « échantillon d'apprentissage » permet (en utilisant un logiciel statistique ayant intégré un certain algorithme) de calculer une estimation de a, ainsi que de b. Admettons que a = 2 et b= -8, l'équation de prédiction aura, dans ce cas, la forme simple suivante : $2 \cdot P - 8 = \hat{C}$, ce qui indique que pour toute personne ayant un score de p=5 au test, on peut « prédire » que sa note d'examen $\hat{C}$ sera 2 avec un maximum de probabilité.

Cependant, un calcul de « prédiction » réellement utile nécessite de construire un *intervalle de confiance* ayant par exemple 95 chances sur 100 de contenir la note réelle que l'individu obtiendra à l'examen. Pour ce faire il est nécessaire de connaître l'écart-type des résidus ou, en d'autres termes, la dispersion des erreurs. Celle-ci dépend de la qualité de la prédiction effectuée par le modèle. bien entendu, le modèle est d'autant plus précis (fiable) que la corrélation entre $\hat{C}$ et C est haute dans l'échantillon d'apprentissage.

Par exemple, si l'on veut prédire la note à un examen de gymnastique à l'aide d'un test de raisonnement, on risque de construire un modèle totalement inefficace, r étant quasi

nul. Au contraire, la note à un examen de mathématiques peut être assez bien prédite grâce à un modèle incluant un test de raisonnement.

Le coefficient r est donc une bonne mesure de la qualité de la prédiction ; s'il vaut 1 (ou -1) la prédiction est parfaite ($\hat{C} = C = P$). S'il vaut 0, la prédiction ne vaut pas mieux que celle du hasard ($\hat{C}$ orthogonal à C). En psychologie, on considère en général qu'une corrélation de plus de .50 est déjà acceptable, mais ce seuil dépend du domaine (*cf.* Gendre, 1977, p.78). Certains auteurs préfèrent interpréter le carré de r, à savoir r^2 *qui équivaut au % de variance commune entre le prédicteur et le critère* (% de variance de l'un « expliquée » par l'autre). Cette valeur est aussi parfois appelée : *coefficient de détermination*. Une corrélation de .50 équivaut à une variance commune expliquée de 25%, soit le quart de la variance totale du critère (il ne faut pas être trop exigeant en psychologie !). Une corrélation de .80 est jugée excellente et des valeurs supérieures sont parfois jugées suspectes...

- L'équation de régression est plus simple si l'on standardise les variables ! En effet, : si la *combinaison linéaire* $\hat{C} = a \cdot P + b$ et le critère C sont en corrélation maximum r (avec a et b « bien choisis »), on peut montrer que $z\hat{C} = r \cdot zP$, donc en fait :

  si on standardise les variables prédicteur *et* critère*, l'équation de prédiction se simplifie car le coefficient de régression est simplement r (pente de la droite de régression), et la constante b disparaît (la droite passe par l'origine).*

Si l'on reprend l'exemple ci-dessus, en admettant que la corrélation entre P et C soit de .50, et en admettant que le score z de P (zP) soit 0.34, alors le score au critère $z\hat{C}$ « note prédite d'examen, standardisée » vaut : .50 · .34 = .17 en score z. Il ne reste alors plus qu'à transformer ce score z en score brut pour retrouver la métrique initiale (en le multipliant par l'écart-type du critère et en ajoutant la moyenne). Mais encore une fois, cette valeur ne suffit pas, il faut calculer un intervalle de confiance autour de la valeur .17 prédite par le modèle.

- *La théorie des erreurs de prédiction*

Le concept de corrélation est évoqué chaque fois que l'on s'interroge au sujet de la force des liens pouvant exister entre (au moins) deux séries d'observations P et C prélevées sur un échantillon supposé tiré aléatoirement d'une population dans laquelle les distributions des variables P et C sont supposées normales. La valeur de cet indice varie entre -1 et +1 et sa formule a été déjà développée dans le cadre des tests d'ajustement à une corrélation théorique.

- Le *carré d'une corrélation* exprime la part de variance commune propre à deux variables. En effet, si on raisonne en scores z, la variance de deux variables C et P est toujours 1 et l'équation de régression liant C à P s'écrit : $z\hat{C} = r \cdot zP$.

Donc, la variance de $z\hat{C}$ = var(r · zP) = r^2var(zP) = $\mathbf{r^2}$, car var(zP) = 1.

La variance des $z\hat{C}$ s'interprète donc comme la variance de zC « expliquée » par la variation du prédicteur zP, et par conséquent : le rapport : r^2 /var (zC) = r^2 /1 = r^2 exprime le *% de variance du critère « expliquée » par le prédicteur*.

- Si la variance de $z\hat{C} = r^2$, on peut trouver la valeur de la variance de l'erreur, ce qui nous permettra ensuite de construire des intervalles de confiance. En effet, les variances étant additives, on peut décomposer la variance du critère en une partie « expliquée » et une autre partie « résiduelle ».

  - La partie *expliquée* étant la variance des scores prédits, qui s'interprète comme l'information commune aux deux variables P et C.

  - La *variance* résiduelle ou « d'erreur » qui représente la partie imprédictible du critère.

La décomposition (une *analyse* au sens propre) de la variance totale du critère en ces deux parties, s'écrit : Var(zC) = Var($z\hat{C}$) + Var(zC- $z\hat{C}$).

Comme Var(zC) = 1, et comme vu ci-dessus : Var($z\hat{C}$) = r^2, on en tire que la variance des zrésidus : Var (zC- $z\hat{C}$) vaut : **1- r^2**. En extrayant la racine carrée, on trouve finalement l'écart-type des zrésidus qui vaut : $\sqrt{1-r^2}$

Les résidus(tant standardisés que bruts) étant centrés en zéro, on retrouve l'écart-type des résidus bruts en les multipliant par l'écart-type du critère :

$$s_{\hat{C}-C} = s_C \cdot \sqrt{1-r^2}$$

Pour un échantillon assez grand, on peut admettre que les erreurs se distribuent normalement autour du score prédit individuel. Un intervalle de confiance pour 100·(1-α)% autour d'un score brut prédit $\hat{C}$ est donc délimité par les bornes suivantes :

- Borne supérieure : $\hat{C}_{sup} = \hat{C} + \left[u_{\left[1-\frac{\alpha}{2}\right]} \cdot s_C \cdot \sqrt{1-r^2} \right]$

- Borne inférieure : $\hat{C}_{inf} = \hat{C} - \left[u_{\left[1-\frac{\alpha}{2}\right]} \cdot s_C \cdot \sqrt{1-r^2} \right]$

Si l'on veut construire un intervalle à 95%, $u_{(1-\alpha/2)}$ est le centile .975 de la distribution normale standardisée (table), égal à 1.96.

- On a donc pu construire, pour chaque individu, un intervalle de confiance ayant 95 chances sur 100 de contenir le score C au critère, au cas où cette information devenait disponible. Cette méthode permet aussi de prévoir que sur 100 scores du critère effectivement observés, 95 d'entre eux seront compris dans l'intervalle défini autour du $\hat{C}$ prédit par le modèle, alors que 5 d'entre eux seront en dehors.

En pratique : **la régression linéaire simple,** ce qu'il faut savoir,
ce qu'il faut faire calculer par un logiciel,
et ce qu'il faut calculer soi-même.

- *Précautions d'usage* : Pour construire et utiliser un modèle de régression, il est recommandé de vérifier les conditions suivantes :
  - un trop petit échantillon ne fournira pas de bonnes estimations des paramètres du modèle, mieux vaut disposer d'au moins 100 personnes pour l'échantillon d'apprentissage ;
  - les distributions des variables utilisées devraient être préalablement testées du point de vue de leur normalité ;
  - un diagramme de dispersion des données devrait confirmer l'idée d'un lien *linéaire* entre les variables,
  - le coefficient de corrélation empirique calculé sur les données ne doit pas seulement être significatif, mais il doit aussi être égal à une valeur considérée habituellement comme « forte » dans le domaine considéré. Aucune technique statistique ne permet de juger de la valeur et de l'intérêt heuristique et pratique d'un coefficient de corrélation. Le test de la corrélation permet seulement de rejeter ou non l'hypothèse de l'indépendance des variables dans la population parente. On peut aussi se baser sur le carré de r qui donne la part de variance commune, mais encore une fois cette valeur doit être rapportée à ce qu'on observe habituellement dans le domaine.
- *Données* : deux variables X et Y, mesurées sur n sujets, satisfaisant à la condition de normalité. X est déclarée *prédicteur* et notée P, Y est déclarée *critère*, et notée C (pour des raisons de conformité avec la théorie qui précède !).
- *Premiers calculs* :
  - L'ordinateur calcule m_P et m_C, s_P et s_C, moyennes et écart-types empiriques. (Aussi calculables à la main, calculette, EXCEL, etc.)
  - On obtient aussi r et son carré ; la p-value pour r est donnée dans la procédure SPSS *corrélation*, mais non dans *régression.*
  - Tous les logiciels statistiques calculent encore la pente a et la constante b (*unstandardized coefficients*).
- *Tests de signification pour r* :

- La significativité statistique du coefficient r peut être testée en utilisant la formule de Fisher et en consultant la table des valeurs critiques pour r).
- SPSS fournit un test basé sur le rapport :
  « Mean Square Regression / Mean Square Residual » qui suit une loi de F à 1 et n-2 degrés de liberté. On peut se contenter d'examiner la p-value de F, si elle est inférieure à 0.05, r est significatif[22].
- SPSS fournit encore un autre test basé sur la pente a de la droite de régression qui est donnée en même temps que b, la constante. Le test consiste en un test d'ajustement de la valeur de la pente à la valeur théorique zéro correspondant à l'hypothèse nulle d'indépendance. La valeur t calculée par SPSS est donnée par le quotient de la pente par son erreur standard.
- Dans la colonne *standardized coefficients* on trouve la pente de la droite de regression en scores z, soit simplement la corrélation déjà apparue plus haut. Le test concerne toujours l'hypothèse d'indépendance qui peut aussi se traduire par H_o : la pente de la droite de régression est nulle.
- Note : les trois tests précédents, que ce soit celui sur r, sur les carrés moyens ou celui de la pente sont rigoureusement équivalents et aboutissent au même résultat. C'est pourquoi il suffit d'en considérer un seul !

• *Construction d'un modèle* : si les conditions de base sont remplies *et* si r est significatif *et* assez élevé, on peut envisager de construire un modèle de régression destiné à « prédire » des scores $\hat{C}$ pour des individus dont on connaît le score à P, mais pas encore celui qu'ils obtiendront à C.

- En utilisant les coefficients a et b donnés par le logiciel, on construit facilement l'équation prédictrice, valable pour les scores bruts.
- Si l'on n'a pas les coefficients a et b, il faut construire une équation en scores z, et utiliser r, cela demande un peu plus d'efforts, car il faut ensuite tout reconvertir scores bruts.
- SPSS et la plupart des logiciels statistiques calculent en un clin d'oeil tous les scores prédits. Si l'on ne dispose pas de logiciel spécialisé, EXCEL fait aussi l'affaire, mais il faut passer par les scores z.

• *Calcul d'intervalles de confiance individuels :*

- Le plus commode est d'obtenir directement les bornes des intervalles de confiance en scores bruts (les scores z sont, pour leur part, plus pratiques pour raisonner…). Il faut donc calculer l'écart-type des résidus bruts.

22. Les carrés moyens calculés par SPSS sont ceux obtenus à partir des scores prédits (« regression ») et des résidus (« residuals »). La compréhension de ce test nécessite celle de l'analyse de variance et du test de F : les carrés moyens « regression » et « résidual » sont considérés comme deux estimations de la variance du critère dans la population. Si cette hypothèse est vraie, leur rapport F doit être compris dans certaines limites, d'où le test.

- Sans logiciel sophistiqué, mais avec EXCEL, on calcule facilement la quantité : $\sqrt{1-r^2}$ que l'on multiplie avec s_C , écart-type du critère. On obtient ainsi l'écart-type des résidus recherché : $s_{\hat{C}-C}$.
- SPSS calcule par défaut une valeur corrigée de cet écart-type sous l'appellation : *standard error of the estimate* (version SPSS 11) ou simplement *standard error*, sous : *model summary*. En français on l'appelle souvent : *erreur-type* sur un score prédit individuel. Sa valeur se trouve en divisant par n-2 (et non par n-1) la somme des carrés des résidus.
- Connaissant $s_{\hat{C}-C}$ ou mieux : l'erreur-type, on calcule facilement les bornes de tout intervalle de confiance pour C, construit autour de la valeur prédite $\hat{C}$. L'intervalle de confiance gaussien à 95% autour de $\hat{C}$ est borné par : $\hat{C} \pm 1.96 \cdot s_{\hat{C}-C}$. Si l'on veut s'approcher au plus près des résultats calculés par SPSS, il faut remplacer $s_{\hat{C}-C}$ par l'erreur-type fournie par le logiciel.
- SPSS calcule ces bornes pour tout individu, mais utilise une loi de distribution des erreurs différente (t au lieu de la loi gaussienne). Pour des échantillons petits, les valeurs de SPSS peuvent être légèrement différentes que celles calculées par la méthode exposée ici.

Exemple :

Voici les données correspondant à deux tests de raisonnement passés à 27 personnes :

TABLEAU 10. Test RGC-20 (Prédicteur ;P)

16	*9*	*14*	*14*	*10*	*11*	*10*	*12*	*9*	*13*	*9*	*9*	*17*	*12*
10	*10*	*12*	*1*	*1*	*10*	*9*	*10*	*13*	*15*	*10*	*12*	*5*	

TABLEAU 11. Test B53 (Critère ;C)

18	*16*	*10*	*18*	*4*	*8*	*10*	*7*	*9*	*8*	*10*	*16*	*20*	*15*
7	*16*	*15*	*6*	*6*	*15*	*8*	*11*	*20*	*16*	*16*	*13*	*7*	

Calculs avec SPSS ou tout autre logiciel spécialisé de statistique :

- La commande *regression-linear* fournit : R (qui devrait être en minuscules pour la régression simple) = .592 / r^2 = .35. / *adjusted R square* qui sert à calculer l'erreur-type : / *Std error of the estimate* = 3.912 est l'erreur-type (écart-type des résidus) calculé à l'aide de *adjusted R square*.
- La table *ANOVA* fournit les sommes de carrés, les degrés de liberté correspondants, les carrés moyens pour les scores prédits, ainsi que pour les résidus. Leur rapport

F (ici = 13.472) est accompagné de sa p-value (Sig. = .001). Des valeurs inférieures à 0.05 indiquent que la régression explique mieux le critère que le hasard. Ce test remplace le test direct de Fisher sur r.

- La table *coefficients* donne les paramètres de la droite de régression et le test sur la pente. Mais il est pratiquement inutile d'écrire l'équation, car : si l'on a pris soin de cocher les cases *unstandardized predicted values* et *prediction intervals individual* dans l'option *save*, on obtient d'un coup les scores prédits et leur intervalle de confiance.

- Tout nouveau score peut être ajouté en bas de la colonne des scores du prédicteur, en faisant « tourner » encore une fois la commande, on obtient son score prédit et l'intervalle de confiance associé. Idem pour tout autre nouveau score.

Calculs avec EXCEL, première variante :

- Les données des deux variables doivent être disposées en colonnes.

- Quoique ce logiciel ne soit pas vraiment fait pour ce genre de calcul, il est possible de calculer toutes les statistiques nécessaires (r, r^2, erreur-type, F, pente et constante) à l'aide la fonction DROITEREG(col.C;col.P;vrai;vrai), mais cette option n'est pas vraiment très accessible (il faut l'entrer en écriture matricielle), elle est peu pratique et ne donne pas de résultats de test sur r. Ce dernier s'obtient en examinant la valeur de F dans une table, ou en transformant r selon la formule de Fisher.

Calculs avec EXCEL, seconde variante :

- Les fonctions élémentaires d'EXCEL permettent de calculer les moyennes, écart-types et coefficient de corrélation r. Par contre les paramètres *pente* et *constante* de la droite de régression ne sont pas calculables simplement. De plus, le test de r doit être effectué à l'aide de la formule de Fisher et de la loi de Student (accessible dans EXCEL).

- Il faut donc passer par les scores z, on standardise le prédicteur à l'aide de sa moyenne (m_P) et de son écart-type (s_P).

- On construit (formule) le modèle en scores z et on crée une colonne de scores z prédits ($z\hat{C}$), que l'on reconvertit dans la métrique du critère en les multipliant par s_C et en leur ajoutant m_C. On obtient donc les $\hat{C}$, autour desquels il faut construire un intervalle de confiance.

- On calcule l'écart-type des résidus bruts par la formule $s_{\hat{C}-C} = s_C \cdot \sqrt{1-r^2}$. Si l'on est perfectionniste, on peut calculer l'erreur-type exacte en calculant les résidus bruts puis en calculant leur écart-type corrigé en divisant par (n-2). Pour des

grands échantillons (n plus grand que 100), cette différence n'a pas grande importance.

- Dans cet exemple, $\sqrt{1-r^2} = .81$, $s_C = \dot{4},76$ et l'écart-type des résidus bruts vaut donc : $4,76 \cdot 0,81 = 3,85$. SPSS donne la valeur exacte : 3.91 (erreur type de prédiction)

- Pour un nouveau score observé 16 à la variable P, la valeur $\hat{C}$ prédite est 16.20, et les bornes de l'intervalle de confiance gaussien construit autour de cette valeur, et ayant 95 chances sur 100 de contenir la vrai valeur C sont : 16.20 $\pm (1,96 \cdot 3,85)$ = **23.7** et **8.67**

- L'intervalle calculé par SPSS, utilisant l'erreur type de 3.91 associée à une distribution de Student, est un peu plus large.

F.2. Cas 2 : *Les modèles de régression linéaire multiple*

Dans ce cas, on a toujours un seul critère C, mais on dispose de plusieurs prédicteurs P_i pour l'approcher au mieux. Un *modèle linéaire prédictif multiple* est une équation du premier degré définissant une combinaison linéaire qui s'écrit : $\hat{C} = B_1 \cdot P_1 + B_2 \cdot P_2 + \ldots + K$.

Comme dans le cas précédent, on collecte un échantillon d'apprentissage (*learning sample)* dans lequel on mesure les P_i et C. Résoudre un problème de prédiction consiste à trouver les meilleurs coefficients B_i ainsi que la constante K, tels que $\hat{C}$ et C soient en corrélation maximum. Les logiciels modernes permettent en général de trouver les nombres nécessaires en quelques secondes.

Le problème de la qualité de la prédiction se pose à nouveau ! Il est intuitivement évident que plus les prédicteurs sont globalement liés au critère, plus la prédiction sera précise. Ce lien est mesuré par la corrélation entre $\hat{C}$ et C, mais comme $\hat{C}$ n'est pas une variable mesurée, mais une variable *construite* (par combinaison linéaire des P_i), on appelle ce coefficient *corrélation multiple* et on le note par convention R. Toujours par convention, les coefficients B_i sont appelés « poids B ». De même que pour le cas des modèles simples, plus R est proche de 1 (ou -1), plus la prédiction est précise, si R est nul, elle est impossible.

On peut aussi *simplifier l'équation* de régression/prédiction multiple en standardisant les variables P_i en zP_i – et C en zC. Dans ce cas, on construit une combinaison linéaire de variables en scores z : $z\hat{C} = \beta_1 \cdot zP_1 + \beta_2 \cdot zP_2 + \ldots + \beta_n \cdot zP_n$ avec $z\hat{C}$ et zC en

corrélation multiple R maximum. Les coefficient β sont appelés dans ce cas les « poids bêtas » qui représentent des *corrélations partielles* entre chaque prédicteur et le critère.

On peut ainsi par exemple chercher à prédire la note à un examen en fonction de plusieurs prédicteurs, mais le problème est souvent de *savoir quel est le meilleur modèle* ? Ce problème important a occupé la carrière de plus d'un auteur. On peut en effet poser la question de l'économie : les prédicteurs étant souvent coûteux à mesurer (temps de passation, etc) et il serait utile de savoir lesquels sont les plus utiles, et lesquels on peut laisser tomber sans trop diminuer la qualité de la prédiction. Ce problème est d'autant plus délicat que très souvent, les prédicteurs sont liés entre eux et qu'il devient difficile d'évaluer l'apport propre de chacun d'eux.

- *Application :* *Un exemple d'application des modèles prédictifs multiples : les tests fonctionnels*

La principale caractéristiques des tests fonctionnels, et qui les distingue de tous les autres tests construits jusqu'à ce jour, est que *les items sont caractérisés* selon un certain nombre de dimensions. Demander à une personne de noter des items selon l'attirance qu'ils exercent sur elle, revient en fait à mesurer son attirance « fondamentale » pour les dimensions sous-jacentes qui sont précisément ces caractéristiques. Cette « attirance » est mesurée par le biais de corrélations entre le vecteur de réponses du sujet avec les n dimensions descriptives des items. Ces corrélations sont ensuite standardisées sur un groupe de sujets et représentent les scores de la personne à des dimensions psychologiques communes aux items *et* aux sujets.

Cette situation peut être modélisée dans le cadre des modèles prédictifs. Le parallèle n'est pas évident, c'est pourquoi nous allons le détailler de la manière suivante :

Dans ce qui suit, il est important de noter que les expressions suivantes sont équivalentes, au niveau d'interprétation près :

- Caractéristiques des items = Echelles ou dimensions fondamentales = prédicteurs du vecteur de réponses
- Corrélations entre caractéristiques et réponses du sujet = scores bruts aux échelles fondamentales = vecteur de stratégie du sujet = pondérations du modèle prédictif des réponses.

Les caractéristiques des items peuvent être associées à des prédicteurs (les P_i) permettant de « prédire » la variable « réponses du sujet » qui joue le rôle de critère (C).

Les attirances du sujet pour les dimensions fondamentales sont mesurées par des corrélations entre C et les P_i. Ces corrélations (standardisées en scores G) des réponses avec les caractéristiques des items, sont interprétées comme des « scores aux dimensions fondamentales », que l'on identifie à des échelles de mesure psychologiques.

On retiendra que *les caractéristiques des items sont en scores z et orthogonales* par construction. L'équation (modèle) prédictive de C s'écrit donc (*cf.* plus haut), en scores z : $z\hat{C} = \beta_1 \cdot zP_1 + \beta_2 \cdot zP_2 + \ldots + \beta_n \cdot zP_n$ où les zP_i sont les caractéristiques des items (standardisées par construction), $z\hat{C}$ les réponses *prédites* du sujet (en scores z), et les β sont les poids bêtas ou coefficients de corrélation *partiels* entre les prédicteurs et le critère.

Or, les zP_i sont orthogonaux par construction ! Il s'en suit que dans l'équation ci-dessus, les bêtas ne sont pas des corrélations partielles, mais des corrélations « normales », égales à celles déjà mesurées ci-dessus.

Les scores bruts d'un sujet aux échelles fondamentales sont donc les coefficients d'un modèle de régression permettant de prédire les réponses *qu'il aurait (!) données*, s'il avait appliqué une stratégie constante tout au long du test.

Par conséquent, on peut également dire que les scores bruts aux échelles fondamentales (qui sont les corrélations entre ses réponses et les caractéristiques des items) sont aussi les pondérations qu'il applique implicitement aux caractéristiques des items, chaque fois qu'il choisit une réponse, quel que soit l'item. C'est pourquoi ce jeu de pondérations ou scores bruts, unique pour chaque individu, a été appelé le « vecteur de stratégie implicite du sujet ».

On peut alors profiter de tous les bénéfices secondaires du modèle : si les P_i sont en scores z *et* standardisés, alors la somme des carrés des poids bêtas équivaut au carré de la corrélation multiple R^2 (on admettra ce fait sans discussion…). Comme, enfin, ces poids ou pondérations sont les scores bruts aux échelles fondamentales (que l'on a calculés par corrélation), alors *la corrélation multiple entre les prédicteurs et le critère (R) est la racine carrée de la somme des carrés des scores bruts* ! Ce nombre est appelé la *cohérence* des réponses et mesure quelque chose de l'ordre de l'adéquation du test (et de son modèle psychométrique sous-jacent) à la personne.

En effet, si une personne répond au hasard, c'est-à-dire ne tient pas du compte des dimensions sous-jacentes, sa cohérence sera nulle, autrement dit : ses réponses sont totalement imprédictibles à l'aide des caractéristiques des items ! En revanche, une personne très à l'aise pour répondre, sensible aux dimensions sous-jacentes, aura une cohérence élevée, ce qui signifie que ses réponses seront très faciles à « prédire » à condition de disposer de son vecteur de stratégie implicite, c'est à dire du jeu de pondérations nécessaire à la construction de l'équation de prédiction des réponses. On notera par ailleurs que le modèle prédictif ainsi construit permet de prédire la réponse à n'importe quel item, à condition qu'il soit caractérisé dans les mêmes dimensions, *même s'il n'appartient pas au test…*

Une autre possibilité d'exploiter le modèle consiste à étudier la différence entre zC et $z\hat{C}$, c'est-à-dire le résidu ou « erreur » de prédiction. Calculer la différence zC - $z\hat{C}$

revient à mettre en évidence les items pour lesquels le modèle se trompe le plus lourdement, par rapport à la réponse que la personne a donnée en réalité. Ces items particuliers sont appelés *singularités* et peuvent être de deux sortes.

- Les items que la personne a notés beaucoup plus haut que ce qui est prédit pas le modèle, et que l'on appelle les *sur-estimés* ;

- et les items notés beaucoup plus bas, appelés *sous-estimés*.

Du point de vue technique, les items sur – et sous-estimés s'isolent en standardisant la différence entre zC et $z\hat{C}$, puis en reportant les items pour lesquels cette différence dépasse un certain seuil, par exemple deux écart-types.

Plus généralement, une *matrice de corrélations* exprime l'information commune, globale, véhiculée par un jeu de p variables numériques. Il existe diverses méthodes pour « structurer » cette information :

- Des méthode purement descriptives (n'impliquant aucune analyse de la variance totale en une part explicable et une part résiduelle, comme en régression) dites « *analyses en clusters* » basée sur le regroupement de variables selon leur « proximités », les distances utilisées pouvant être soit de corrélations (plus elles sont hautes, plus la distance est petite), soit des distances euclidiennes ou encore d'autres types de distances.

- Des méthodes analytiques ayant pour but de construire un *modèle réduit* permettant d'expliquer une part de la variance totale. Si ce modèle est déduit des seules données, on est en présence d'analyses factorielles *descriptives* ou *exploratoires*, et si le modèle est théorique et externe aux données, on est en présence d'analyses *confirmatoires*. Les méthodes de ce type sont couramment appelées *analyse factorielle* si les objectifs sont *confirmatoires (analyse de pistes causales, LISREL, etc.)*, ou *analyse en composantes principales* (ACP) si l'on se limite à une perspective descriptive (qui est pratiquement toujours celle des psychologues actuels).

G. Structuration de données

G.1. *L'analyse en clusters ou analyse typologique*

Avant d'aborder des techniques plus complexes, nous allons exposer une procédure manuelle très simple qui permet de comprendre la structure des liens existant dans un jeu de variables : il s'agit de l'analyse typologique ou *analyse factorielle du pauvre*. Peu de gens la connaissent encore, c'est pourquoi nous l'exposons ici, ne serait-ce que pour pailler à l'absence possible d'ordinateurs, dans une situation ou une autre.

- *Analyse typologique à partir d'une matrice de distances*

On part en principe d'une matrice de distances, mais celles-ci sont parfois plus difficiles à calculer que des corrélations, c'est pourquoi nous utiliserons ce dernier cas de figure, en notant qu'il est toutefois bien commode de calculer les corrélations avec une machine.

1. Soit un jeu de p variables $v_1 \dots v_p$ et leurs corrélations, dresser la matrice de toutes les corrélations (symétrique) avec les unités dans la diagonale.

2. Dans chaque **colonne**, identifier et souligner la corrélation la plus élevée.

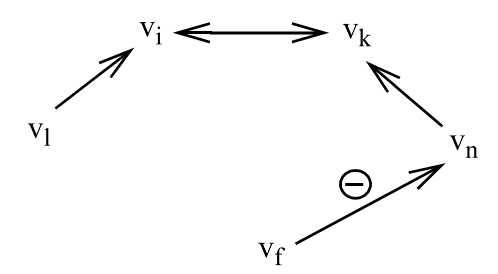
3. Identifier parmi ces dernières, la corrélation la plus élevée de la matrice, elle détermine le noyau du premier cluster, noter sur un papier brouillon :

   $$v_i \longleftrightarrow v_k \quad \text{ou} \quad v_i \overset{\ominus}{\longleftrightarrow} v_k$$

   le signe négatif s'inscrit si la corrélation est négative, mais il est aussi possible d'inverser la variable en changeant son nom (par exemple : sentiment d'infériorité –> sentiment de supériorité).

4. Dans la ligne de v_i, chercher une autre corrélation soulignée (hormis la précédente), elle détermine la seconde variable qui a sa relation la plus forte avec v_i, lier celle-ci au cluster à l'aide d'une flèche qui va de la nouvelle variable vers v_i. Regarder ensuite dans la ligne de cette variable s'il y a une corrélation soulignée, si oui, lier cette nouvelle variable de la même manière et *etc*. jusqu'à ce qu'il n'y aie plus de corrélation soulignée dans la ligne de la dernière variable liée au cluster par v_i.

5. Opérer exactement de la même manière avec v_k, et lier les éventuelles variables au cluster jusqu'à ce qu'il n'y aie plus de corrélation soulignée. On obtient un graphe du genre :



   Le premier cluster est ainsi constitué, les angles des flèches n'ont pour l'instant pas d'importance et on s'efforce de dessiner des flèches d'autant plus longues que les corrélations sont faibles. *Le sens des flèches signifie toujours que la variable « origine » a ses relations les plus fortes avec les variables situées à la « pointe »*.

6. On fait abstraction des lignes et colonnes constituant le premier cluster et on identifie la corrélation la plus haute dans la matrice résiduelle. Le deuxième cluster est constitué de la même manière, puis le troisième, et les suivants s'il y a en a.

7. Les clusters dessinés sur ce premier schéma vont ensuite être liés entre eux en représentant les secondes liaisons les plus fortes entre variables, on les représentera par des flèches de couleur différentes.

8. Pour ce faire, souligner dans chaque colonne la *deuxième* corrélation la plus forte et procéder colonne par colonne : on liera les deux variables entretenant cette corrélation par une flèche allant de la variable figurant dans la première colonne à la variable figurant dans la ligne correspondante (… à la seconde corrélation la plus élevée soulignée dans cette première colonne…). Et ainsi de suite. On voit que certaines flèches vont d'un cluster à un autre, ce qui permet de mieux fixer leurs positions respectives.

9. On peut encore souligner la troisième corrélation la plus forte dans chaque colonne et représenter cette troisième relation par une flèche d'une autre couleur, en suivant les mêmes règles que précédemment, pour ce qui est du sens des flèches.

10. Il est alors temps de recopier les clusters de manière à les placer de manière harmonieuse et aussi claire que possible… c'est possible !

11. On peut encore identifier les *prototypes*, ce sont, pour chaque cluster, la variable qui entretient les corrélations les plus fortes avec toutes les autres. Il suffit donc de calculer la somme des carrés des corrélations dans chacune des colonnes des variables constituant un cluster pour l'identifier : il s'agit de la variable totalisant la somme la plus élevée. Il y a donc autant de prototypes que de clusters, mais le prototype n'est pas nécessairement une des variables du « noyau ». Une analyse de second ordre est possible en reconstituant une nouvelle matrice de corrélations en ne prenant que les prototypes, on obtient ainsi une « superstructure » parfois plus claire, mais aussi plus réduite, que celle de niveau inférieur.

12. Enfin, l'analogie avec l'analyse factorielle peut être plus poussée : chacun des prototypes représente un facteur, et les corrélations des variables du cluster avec lui-même sont comme les saturations de ces variables dans leur facteur. On aboutit ainsi à une sorte de structure dont les éléments sont obliques (non orthogonaux), et dont on peut vérifier la validité en soumettant le jeu de variables à une analyse en composantes principales ordinaire : les résultats des deux méthodes ne sont souvent pas très différents.

G.2. *Les modèles factoriels*

Contrairement aux modèles de prédiction, qui s'expriment la plupart du temps sous forme d'une équation, parfois d'une matrice de probabilités (chaînes de Markov, par exemple), les modèles factoriels sont des systèmes de repères ou bases (au sens algébrique ou géomètrique du terme) dans lesquelles on cherche à représenter des variables ou des individus. On rencontre aussi fréquemment le terme de « structure » qui désigne aussi un

système de repères, à condition que celui-ci obéisse un à un certain nombre de contraintes liées à la « simplicité ». Dans le domaine de la recherche et de la modélisation en sciences sociales, ces contraintes de simplicité s'expriment généralement par trois conditions :

- Les éléments de la structure doivent être facilement interprétables et si possible non redondants,
- Ces éléments doivent être en nombre minimum, mais doivent représenter la majeure partie de l'information contenue dans un ensemble de données redondantes et difficiles à interpréter,
- Chaque élément de cette structure porte une certaine quantité d'information, indépendamment des autres, et cela dans un ordre hiérarchique : le premier élément porte le maximum d'information, le dernier le moins.

Dans un vocabulaire plus technique, une structure factorielle permet de modéliser, dans le but de la clarifier, la structure complexe des inter-relations entre plusieurs variables « originales » redondantes.

Un élément de la structure factorielle est une variable latente (un facteur) supposée expliquer la variance commune de plusieurs variables intercorrélées. Ainsi plusieurs groupes de variables intercorrélées peuvent être remplacées par autant de facteurs uniques représentant chacun, à lui seul, l'essentiel de l'information véhiculée par un groupe de variables liées. Ces facteurs devant être, si possible, indépendants les uns des autres. Historiquement, les premiers facteurs ont été construits par Spearman (début du 20e siècle) dont l'objectif était de représenter la part de variance commune observée entre plusieurs tests. Cette part prépondérante fut associée à de l'intelligence, et le facteur supposé la mesurer fut baptisé « facteur g ». Ainsi, l'intelligence devenait le facteur principal permettant d'expliquer la part la plus importante de la variation commune de plusieurs tests. D'autres facteurs plus spécifiques furent ensuite dégagés : raisonnement (R), spatial (S), verbal (V), etc. De là vint sans doute l'expression « Analyse en Composante Principale » (ACP) qui désigne la technique de mise en évidence de structures factorielles.

Une structure factorielle peut être *donnée par l'expérience*, on parle alors d'ACP *exploratoire* ; mais elle peut aussi être déterminée par une théorie et sa structure soumise à des contraintes décidées *a priori*, on parle alors d'ACP *confirmatoire*. Le terme d'« analyse factorielle » est actuellement un peu confus et désigne des techniques nombreuses et différentes. On lui reconnaît toutefois un usage générique, surtout dans les logiciels qui proposent l'option générale *factor analysis*, qui renferme alors des variantes appelées ACP, confirmatory analysis, etc. En bref, les psychologues se comprennent tout de même assez bien lorsqu'ils parlent d'analyse factorielle, mais il faut reconnaître que les statisticiens sont plus pointilleux sur les termes utilisés.

- *Du point de vue technique*, on peut construire autant de facteurs qu'il y a de variables originales. Chaque facteur est une combinaison linéaire de celles-ci, au sens des régressions multiples. Le choix des coefficients est évidemment un problème ardu étant donné les contraintes imposées : les facteurs doivent tous être orthogonaux entre eux, et leurs variances (=l'information) doivent être décroissantes. À vrai dire, les mathématiciens avaient déjà la solution à ce problème dès le 19e siècle, bien avant que les statisticiens-psychologues ne posent le problème. Nous nous contenterons ici de dire que ce problème trouve sa solution dans la décomposition spectrale (due à Eckart & Young) de la matrice de corrélations des variables originales.

- Les corrélations entre les variables originales et les facteurs s'appellent les saturations, elles permettent de nommer (interpréter) les facteurs et de leur donner un sens psychologique.

- La corrélation multiple entre les facteurs et chaque variable originale s'appelle la *communalité* (racine carrée de la somme des carrés des saturations en ligne), cet indice permet de savoir à quel point chaque variable est bien représentée par la structure factorielle.

- La corrélation multiple entre les variables originales et chaque facteur s'appelle le *% de variance totale expliqué par chaque facteur* (somme des carrés des saturations en colonne). Cet indice montre la représentativité de chacun des facteurs, il est en relation directe avec la variance de chacun d'eux.

- On cherche en général a construire un modèle optimal, réduit à quelques facteurs, exprimant à eux seuls l'information utile et interprétable. Divers critères (Kaiser, Cattel, etc.) permettent de choisir le nombre de facteurs à retenir.

- *Un cas particulier : orthogonalisation de dimensions descriptives d'items (modèle de mesure fonctionnel)*

Nous avons vu que les items d'un test pouvaient tous être décrits au moyen d'un certain nombre de caractéristiques dont le choix dépend de la méthode utilisée :

- On peut soumette les items à un échantillon de personnes et analyser la structure de leurs interrelations. L'analyse en Composantes Principales dégagera une structure orthogonales dont les éléments peuvent être interprétés. Pour caractériser les items, on peut utiliser les corrélations (saturations) de chacun d'eux avec les facteurs. Ces saturations n'étant pas orthogonales, on doit ensuite les orthogonaliser, conformément aux exigeances décrites dans le chapitre précédent.
- On peut aussi charger un groupe d'experts de caractériser les items selon des caractéristiques *a priori*. Dans ce cas, la moyenne de leurs évaluations à chaque caractéristique constituera une variable, et l'ensemble de ces variables peut être soumis à l'ACP. Le résultat de cette technique sera directement orthogonal.

Dans le cas où les caractéristiques ne sont pas indépendantes (comme dans le premier cas ci-dessus), il est possible de les orthogonaliser, à condition que leurs intercorrélations ne soient pas trop fortes. Orthogonaliser une série de n variables suppose d'effectuer une ACP avec rotations Varimax, en exigeant que la solution comporte autant de facteurs que le nombre de variables de départ. Les saturations permettent d'identifier les facteurs et de vérifier qu'ils correspondent bien aux variables originales (ils se présentent souvent dans un ordre différent). Les caractéristiques d'un test fonctionnel de bonne qualité devraient être construites en combinant les résultats de ces deux méthodes.

APPENDICE :

EXERCICES DE LECTURE DE TABLES ET QUESTIONS DE STATISTIQUE

I. Loi de répartition normale - standardisation

1.1. Si x est le score maximum de X, quel est son percentile ?

Réponse :

1.2. Quel est le percentile du médian ?

Réponse :

1.3. Si moyenne et médian sont confondus, dans quelle proportion partagent-ils la distribution des scores ?

Réponse :

2. *Standardiser* revient à *centrer* et *réduire* une distribution.

2.1. Si m = 100 et s = 15, quel est le score standard (score z) de x = 85 ?

Réponse :

2.2. Si m = 100 et s = 15, le score standard (score z) de x est 1, quel est alors x ?

Réponse :

3. Usage de la loi normale réduite - *cf.* table de u.

3.1. Quel est le percentile du score x = m + s, (répartition supposée normale) ?

Réponse :

3.2. Quel est le percentile du score x = m – s, (répartition supposée normale) ?

Réponse :

3.3. Quelle est la proportion de scores compris dans l'intervalle [m ± s], (id.) ?

Réponse :

3.4. Soit une distribution (supposée normale) de moyenne = 20 et d'écart-type = 6, quel est le pourcentage de scores inférieurs à 15 ?

Réponse :

3.5. Soit une distribution (supposée normale) de moyenne = 60 et d'écart-type = 10, 20% des scores sont supérieurs à ....

Réponse :

3.6. Soit un test d'aptitudes dont les résultats (ou *scores*) sont supposés distribués normalement et arrondis à .5, leur moyenne est 50 et l'écart-type est 10. Supposez que vous deviez sélectionner le 40% d'individus ayant obtenu les meilleurs résultats, quel est le meilleur score *non sélectionné* ?

Réponse :

3.7. Utilisant les mêmes résultats qu'à la question précédente, vous décidez d'« homogénéiser » le groupe d'individus en éliminant le 20% des moins « forts », et le 20% des plus « forts ». Quels sont les scores « critiques » ?

Réponse : inférieurs à .... et supérieurs à ..... (arrondis au dixième près)

4. Certains résultats statistiques peuvent être gravement faussés par la présence de scores (objets, individus, etc.) considérés comme *aberrants* (en anglais : *outliers*). On les reconnaît à leur grande excentricité par rapport à la moyenne. Leur probabilité d'apparition est très faible et on les trouve donc très loin de la moyenne, au delà des queues de la loi normale. En éliminant les cas de ce type, on évite d'intégrer dans les calculs des scores qui résultent souvent d'erreurs de frappe, ce qui évite de biaiser gravement les calculs de moyennes qui auraient des répercussions désastreuses sur tous les autres indices statistiques. Le critère de dépistage des outliers est simple, mais il faut avant tout s'assurer qu'on est bien en présence d'une erreur manifeste, et non d'un individu particulier dont l'élimination ne serait pas réellement justifiée si l'intention de l'étude est de comprendre la réalité dans toute sa diversité. En général, on élimine tous les sujets dont le score n'est pas compris dans l'intervalle [m ± 3s].

4.1. Considérant la distribution de la question précédente (3.6), déterminez les seuils critiques au delà desquels un score peut être considéré comme « aberrant ».

. . . .

4.2. Dans la distribution (ci-dessous en haut de la page suivante, m = 25.2 et s = 8.5), obtenue à partir de scores à un test verbal, identifiez un score « aberrant » et expliquez sa présence.

. . . .

-7	1	.4	.4	.4
1	1	.4	.4	.8
2	1	.4	.4	1.3
3	1	.4	.4	1.8
5	4	1.7	1.8	3.5
7	1	.4	.4	4.0
8	2	.9	.9	4.8
9	1	.4	.4	5.3
10	2	.9	.9	6.2
11	1	.4	.4	6.6
12	1	.4	.4	7.0
13	2	.9	.9	7.9
14	4	1.7	1.8	9.7
15	4	1.7	1.8	11.5
16	4	1.7	1.8	13.2
17	13	5.5	5.7	18.9

II. Distributions échantillonales de moyennes, loi de Student

1. Étudier un caractère *en général* (attribut, dimension, paramètre, etc... le vocabulaire est large et mal spécifié !) revient - en statistique - à chercher à connaître les caractéristiques (moyenne, variance, extrêmes, etc.) de sa distribution dans la population qui est au centre de nos intérêts. Comme cette population ne peut pas être considérée dans son entier, le chercheur doit pratiquement toujours se contenter d'étudier des échantillons partiels, mais si possible représentatifs de cette population. Chercher à connaître le général à partir d'informations collectées sur des réalités particulières, limitées, revient à utiliser la *pensée inductive*, qui en statistique se réalise dans un ensemble de techniques relevant de....
   ....
   ....

2. Si l'on veut que le « passage » (par induction) du particulier au général (ou : de l'observation au modèle) soit pertinent et productif en matière de connaissances, il faut impérativement que les échantillons étudiés remplissent certaines conditions. Ils doivent être avant tout .... de la population « parente », et pour cela doivent en principe être tirés .... D'autre part, ils doivent permettre des estimations fiables et suffisamment précises pour permettre la construction de modèles utilisables. C'est pourquoi les échantillons doivent si possible être « assez grands ».

3. Soit un échantillon comportant n sujets (« cases » en anglais, par ex. dans SPSS) tirés au hasard ; pour un caractère donné, chaque cas donne lieu à une observation consignée généralement sous la forme d'une réponse, on définit ainsi une observable (taille, nom, âge, aptitude au raisonnement, etc). Si les observations sont numériques et ordonnables, on parle d'une variable X (représentant le caractère étudié) qui se « réalise » pour chaque sujet sous la forme d'un « score ». On peut décrire la distribution de ces scores à l'aide des indices statistiques (nombres) usuels, à savoir :
   . . . .

4. Il existe une autre catégorie de variables, à savoir celles qui associent des nombres non pas à des sujets, mais à des échantillons. Ainsi, on peut associer à tout échantillon de taille n tiré toujours de la même population un nombre appelé moyenne. Cette quantité est une variable .... car elle associe un nombre (la moyenne) à tout échantillon de taille fixée n. Ainsi, variance, moyenne, écart-type, etc. sont aussi des variables qui se réalisent en des nombres particuliers, pour un échantillon donné.

5. Certaines variables échantillonnales sont utilisées comme des « variables de décision », celles-ci sont utilisées dans les tests inférentiels que l'on appelle aussi « tests d'hypothèse » ou encore « tests de signification ». Les variables de décision les plus connues sont z, t (Student), « chi carré » (Pearson), r (coefficient de corrélation de Bravais-Pearson), F (de Fisher-Snedecor), etc...

6. Les variables échantillonnales *moyenne*, *variance*, etc., sont aussi utilisées comme des *estimateurs* des paramètres théoriques de la population. Les valeurs numériques observées dans un échantillon sont des estimations de ces valeurs théoriques, inconnaissables. Par exemple la variable échantillonnale moyenne : $M = \Sigma x_i/n$ est l'*estimateur* de la moyenne μ d'une population. Si on tire un échantillon et qu'on en calcule la moyenne m, alors m est une *estimation* de μ. Les estimateurs se représentent par des ...., alors que les estimations sont des .... (= réalisations de l'estimateur pour un échantillon donné = observations).

7. Les caractéristiques de la variable échantillonnale « moyenne » sont entièrement déterminées par le *Théorème Central Limite*. On peut en effet démontrer que la moyenne d'échantillons de taille n tirés aléatoirement d'une population de moyenne μ et d'écart-type σ suit une loi normale de moyenne μ et d'écart-type $\frac{\sigma}{\sqrt{n}}$. Le T.C.L. est très important car il constitue le fondement de l'inférence statistique. De plus, on sait qu'il s'applique même si la distribution originale de X dans la population n'est pas gaussienne ! Un autre très grand intérêt du T.C.L. est qu'*il permet la standardisation de la variable échantillonnale M*. Puisqu'on connaît la moyenne (μ) de M et son écart-type $\frac{\sigma}{\sqrt{n}}$, alors la variable : $z = \frac{M - \mu}{\frac{\sigma}{\sqrt{n}}}$ est gaussienne, de moyenne zéro et d'écart-type 1, et sa distribution correspond à celle du *u* de la table.

   7.1. Soit une population dans laquelle un caractère X est distribué plus ou moins normalement, avec une moyenne μ = 50, et un écart-type
   σ = 20. Quelle est la probabilité qu'une moyenne calculée sur un échantillon de taille 100 dépasse 52 ?.. . . . .

   . . . .

7.2. Dans le même échantillon, quels sont les scores ayant une probabilité totale de 5% d'être dépassés ?. . . . .

. . . .

8. Si on ne connaît pas l'écart-type σ de X dans la population (ce qui est pratiquement toujours le cas !), on est forcé de l'*estimer* grâce à l'estimateur *écart-type de X dans l'échantillon.* L'écart-type théorique inconnu σ sera donc remplacé, pour l'échantillon déterminé, par la valeur de l'estimateur : $\sqrt{\sum_{1}^{n} \frac{(x_i - m)^2}{n-1}} = S$.

M peut donc toujours être standardisée, mais l'utilisation de l'écart-type empirique en lieu et place du théorique *introduit un biais* qui a pour effet que la variable standardisée : $T = \frac{M - \mu}{\frac{S}{\sqrt{n}}}$ n'est plus distribuée normalement. Elle suit par contre une loi assez proche, mais dépendante de la taille de l'échantillon, nommée *loi du t de Student* à [n-1] degrés de liberté.

Notons que lorsque n est grand, le biais dû à l'utilisation de S au lieu de σ perd de son effet et la variable M standardisée suit une loi très proche de la normale.

8.1. Soit une population dans laquelle un caractère X est distribué plus ou moins normalement, avec une moyenne $\mu = 50$ et un écart-type inconnu. Quelle est la probabilité qu'une moyenne calculée sur un échantillon de taille 36 ($S = 24$) dépasse 60 ?. . . . .

. . . .

8.2. Dans le même échantillon, quels sont les scores ayant une probabilité totale de 5% d'être dépassés ?. . . . .

8.3. Même question que 8.1, mais l'échantillon est de taille 100. . . . .

. . . .

8.4. Même question que 8.2, mais l'échantillon est de taille 100. . . . .

. . . .

III. Intervalles de confiance

1. Un intervalle de confiance gaussien à α% définit un intervalle dans lequel une certaine valeur x_i d'une distribution gaussienne a α% de chances de se trouver. Construire un intervalle de confiance revient donc à trouver les bornes supérieures et inférieures de la distribution, au-delà desquelles une valeur x_i n'a que 1-α chances de se trouver. Les intervalles de confiance étant en général symétriques, il suffit donc de trouver le percentile de la distribution correspondant à α. Les intervalles de confiance usuels sont définis pour 95%, ils excluent donc les 2,5% extrêmes de la distribution. Les bornes d'un tel intervalle se trouvent en cherchant le percentile 97.5 de la distribution normale standard, à savoir 1.96. Pour cette distribution très particulière, l'intervalle de confiance s'écrit : [-1.96; 1.96], il est bien entendu centré en zéro.

   On en déduit que pour toute distribution centrée en m et d'écart-type s, l'intervalle de confiance à 95% sera centré en m et borné par $\pm$ 1.96·s.

   1.1. Déterminer un intervalle de confiance à 95% pour une mesure distribuée normalement, dont la moyenne est 50 et l'écart-type 20.

   ....

   ....

   ....

2. Supposons que l'on connaisse la moyenne d'une population (μ), mais non sa variance. Cela n'empêche pas de se demander si la moyenne d'un échantillon de taille n sera compris dans un certain intervalle de confiance à 95% autour de μ. En effet, pour un échantillon de taille n, on sait que la variable échantillonnale M suit une loi de Student à n-1 degrés de liberté et aura une espérance μ et un écart-type de $\frac{S}{\sqrt{n-1}}$ (S étant l'écart-type du caractère dans l'échantillon). Il faut alors chercher le percentile 97.5 de la distribution de $t_{[n-1]}$, qui dépend de n, que l'on peut noter : $t_{1-\alpha/2[n-1]}$. L'intervalle de confiance s'écrit alors :

$$[\mu - t_{1-\alpha/2[n-1]} \cdot \frac{S}{\sqrt{n-1}} \; ; \mu + t_{1-\alpha/2[n-1]} \cdot \frac{S}{\sqrt{n-1}} \;]$$

2.1. Soit un échantillon de taille n = 17, de moyenne = 70 et d'écart-type = 20. Déterminer un intervalle de confiance ayant 95% de chances de contenir la moyenne de la population (c'est le problème inverse du point précédent, mais il se résout de la même façon).

. . . .

. . . .

. . . .

2.2. Même question, mais la taille de l'échantillon est n = 101, puis comparer le résultat avec celui obtenu en utilisant une loi normale au lieu d'une loi de t.

. . . .

. . . .

IV. Estimation d'une variance théorique, variance d'une distribution de moyennes, loi du « Chi-carré »

1. Les formules permettant le calcul des valeurs de t suivant une loi de Student montrent que la variance S^2 d'un échantillon de taille n permet directement d'estimer la variance théorique σ^2 de la population dont il est tiré[23]. Ainsi, la pondération (par leurs effectifs) de plusieurs variances d'échantillons peut constituer une bonne estimation de la variance théorique, même si les échantillons ne sont pas de même taille.

   Il existe une autre manière d'estimer σ^2 qui est basée sur la *variance de la variable échantillonnale des moyennes* (M). Soient les M_i, scores de cette variable M, et μ son espérance. La variance S_M^2 des M_i s'écrit : $\sum_{1}^{p} \frac{(M_i - \mu)^2}{p-1}$ p étant le nombre d'échantillons, tous de même taille n. Dans cette formule, μ est un nombre et les M_i constituent une variable distribuée de manière gaussienne d'après le « Théorème central limite », les différences (M_i - μ) sont donc aussi distribués de manière gaussienne, mais leur carré n'est par contre pas gaussien, il suit une loi dite du « chi carré un » et la somme de ces carrés suit, selon la théorie[24], une loi dite du « chi carré [p-1] » qui se note : $\chi^2_{[p-1]}$. Le terme [p-1] définit ce que l'on appelle les *degrés de*

23. Car l'espérance de S^2 est $[(n-1)/n] \cdot \sigma^2$, c'est-à-dire pratiquement la variance de la population.

24. Une variable constituée par la somme de k carrés de lois normales indépendantes suit une loi dite du chi carré à k-1 degrés de liberté.

liberté de la loi en question, égal au nombre de carrés de différences ajoutés, moins un. Il faut savoir qu'il existe une relation fondamentale entre σ^2 et S_M^2 (variance des moyennes d'un échantillonnage de taille n) qui permet d'estimer la première variance à partir de la seconde, la théorie montre que la quantité :

(1) $\frac{n \cdot S_M^2}{\sigma^2}$ suit une loi χ^2 à p-1 degrés de liberté

La technique dite de l'*analyse de variance* exploite précisément cette double possibilité d'estimer une variance théorique, l'une à partir des simples variances des échantillons, et l'autre à partir de la variance de la variable échantillonnale M des moyennes (σ^2 s'estime par $n_1 \cdot S_M^2$). Le rapport de ces deux variances suit une loi dite de F, qui est tabulée (*cf.* point suivant), ce qui permet de savoir dans quelle mesure l'une des deux estimations excède l'autre.

2. Comme la loi du *t de Student*, la loi du chi carré est donc associée à un certain nombre de degrés de liberté. Il existe des tables permettant de connaître les fractiles des lois du *chi carré* pour divers degrés de liberté :

2.1. La forme d'une distribution (loi) du χ^2 dépend de . . . .

2.2. La moyenne d'une loi $\chi^2_{[23]}$ vaut. . . . .

2.3. Que vaut $\chi^2_{(1-0.05)[1]}$, autrement dit, quel est le percentile 95 de $\chi^2_{[1]}$? . . . .

2.4. Que vaut $\chi^2_{(1-0.05)[2]}$, autrement dit, quel est le percentile 95 de $\chi^2_{[2]}$? . . . .

2.5. Que vaut $\chi^2_{(1-0.01)[1]}$, autrement dit, quel est le percentile 99 de $\chi^2_{[1]}$? . . . .

2.2. Que vaut $\chi^2_{(1-0.01)[2]}$ autrement dit, quel est le percentile 99 de $\chi^2_{[2]}$? . . . .

2.7. Que vaut $\chi^2_{(1-(0,5))[7]}$ autrement dit, quel est le médian de $\chi^2_{[7]}$?.. . . . .

2.8. Quelle est la probabilité de trouver un $\chi^2_{[3]}$ supérieur à 9.348 ?. . . . .

2.9. Quelle est la probabilité de trouver un $\chi^2_{[3]}$ supérieur à 12.838 ?. . . . .

2.10. Quelle est la probabilité de trouver un $\chi^2_{[3]}$ égal à 2.366 ?. . . . .

2.11. Quelle est la probabilité de trouver un $\chi^2_{[3]}$ supérieur à 1.869 et inférieur à 6.25 ? . . . .

V. Distribution de rapports de variances et loi « F » de Fisher

1. Nous avons vu dans les points précédents qu'une variance théorique peut s'estimer de plusieurs manières, en particulier à l'aide des variances d'échantillons, mais aussi à l'aide de la variance des moyennes d'échantillons de taille fixée.
   La théorie statistique s'est aussi intéressée à connaître la *distribution de rapports d'estimations de variances.* Toute variance étant une somme de carrés de différences, sa distribution échantillonnale - pour une taille d'échantillon fixée - suit une loi de chi carré dépendante de la taille de l'échantillon. Par exemple, la variance d'un échantillon de taille n suit une loi $\chi^2_{[n-1]}$, on dit aussi que le nombre de degrés de liberté attaché à la somme de carrés est n-1. Pour une estimation de variance obtenue à partir d'une variance de moyennes, l'estimation [n· S^2_M] est associée au nombre de degrés de liberté p-1, p étant le nombre d'échantillons (groupes) intervenant dans le calcul de la moyenne et n le nombre d'individus d'un échantillon.

   De manière générale, et pour simplifier, on admettra que le quotient de deux estimations de la même variance théorique suit une loi F de Fisher associée aux degrés de liberté des deux estimations.

2. Par exemple, pour un échantillon de taille n_1 et un autre de taille n_2 , le rapport de leurs variances suit une loi de F (la plus grande variance est toujours placée au numérateur !) à (n_1-1) et (n_2-2) degrés de liberté (l'ordre des degrés de liberté dépend de la taille des variances). Ce théorème a déjà été utilisé dans le cadre du test d'homogénéité des variances de deux échantillons indépendants. On écrit dans ce cas : $\frac{S_1^2}{S_2^2}$ Ý

$F_{[(n_1-1);(n_2-1)]}$, ce qui signifie que le rapport des variances empiriques suit une loi de F à (n_1-1) et (n_2-2) degrés de liberté. Ces lois F sont tabulées et, par exemple dans celles de Saporta (p.98), la réalisation F de la variable $\frac{S_1^2}{S_2^2}$ (rapport des variances empiriques) est associée aux indices v_1 et v_2 qui correspondent aux degrés de liberté (n_1-1) et (n_2 - 2) respectivement. Le premier degré de liberté correspondant à la variance empirique la plus grande, placée au numérateur.

2.1. Que vaut $F_{(1-0.05)[20, 3]}$ autrement dit, quel est le percentile 95 de $F_{[20, 3]}$? . . . .

2.2. Quelle est la probabilité de trouver un $F_{[15, 4]}$ supérieur à 14.2 ? . . . .

2.3. On suppose que deux échantillons (n_1 = 20, n_2 = 30) sont tirés d'une même population de variance théorique inconnue pour un caractère X, quelle est la probabilité d'observer un rapport des variances empiriques inférieur à 1.96 ?
(Note : la variance du premier échantillon est supposée plus grande que celle du deuxième)
. . . .

Sources et références

Bavaud, F., Capel, R., Crettaz, F. & Müller, J.-P. (1996). *Guide de l'analyse de données avec SPSS 6*. Genève : Slatkine (épuisé).

Desrosières, A. (1993). *La politique des grands nombres, histoire de la raison statistique*. Paris : La Découverte.

Capel, R., Monod, D. & Müller, J.-P. (1996). Essai sur le rôle des tests d'hypothèse en sciences humaines, rite propitiatoire ou pièce à conviction ? *Actualités psychologiques,* 1, (1), pp. 1-50.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. (2th ed.). Hillsdale NJ : Erlbaum.

Gendre, F. (1976). *L'analyse statistique multivariée*. Genève : Droz.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds). *A handbook for Data Analysis in Behavioral Science – Methodological Issues* (pp. 311-339). Hillsdale : Lawrence Erlbaum.

Fisher, R. (1935). *The design of experiments*. (8th ed. 1966). Edinburgh : Oliver & Boyd.

Howell, D. C. (1998). *Méthodes statistiques en sciences humaines*. Bruxelles : De Boeck.

Huberty, C. J. (1993). *Historical origins of testing practices : the treatment of Fisher versus Neyman-Pearson views in textbooks*. Journal of Experimental Education, 61, (4) 317-333.

Hunter, J. E. (1997). Needed : a ban on the significance test. *Psychological Science*. 8, 3 - 7.

Saporta, G. (1990). *Probabilités, analyse de données et statistique*. Paris : Technip.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals : an integreted alternative method of conducting null hypothesis statistical tests. *Psychological Methods*. 6, (4) 371-386.

Salsburg, D. S. (1985). The religion of statistics as practiced in medical journals. *American Statistician*, 39 (3), 220-223.